

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
18 October 2001 (18.10.2001)

PCT

(10) International Publication Number
WO 01/77125 A2

- (51) International Patent Classification⁷: **C07H**
- (21) International Application Number: **PCT/US01/10708**
- (22) International Filing Date: **4 April 2001 (04.04.2001)**
- (25) Filing Language: **English**
- (26) Publication Language: **English**
- (30) Priority Data:
60/194,361 **5 April 2000 (05.04.2000)** **US**
- (71) Applicant (for all designated States except US): **GENAIS-
SANCE PHARMACEUTICALS, INC.** [US/US]; Five
Science Park, New Haven, CT 06511 (US).
- (72) Inventors; and
- (75) Inventors/Applicants (for US only): **CHOI, Julie, Y.**
[US/US]; 38 Elizabeth Street, West Haven, CT 06516
(US). **KLIEM, Stefanie, E.** [DE/US]; 1298 Hartford Turn-
pike, Apt. 4E, North Haven, CT 06473 (US). **KOSHY,
Beena** [IN/US]; 1298 Hartford Turnpike, Apt. 11B, North
Haven, CT 06473 (US).
- (74) Agents: **SHANER, Sandra, L.** et al.; Genaissance Phar-
maceuticals, Inc., Five Science Park, New Haven, CT
06511 (US).
- (81) Designated States (national): **AE, AG, AL, AM, AT, AU,
AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ,
DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR,
HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR,
LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ,
NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM,
TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.**
- (84) Designated States (regional): **ARIPO** patent (GH, GM,
KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian
patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European
patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE,
IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF,
CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).
- Published:
— without international search report and to be republished
upon receipt of that report
- For two-letter codes and other abbreviations, refer to the "Guid-
ance Notes on Codes and Abbreviations" appearing at the begin-
ning of each regular issue of the PCT Gazette.

(54) Title: **HAPLOTYPES OF THE CCR5 GENE**

(57) Abstract: Novel single nucleotide polymorphisms in the human chemokine (C-C motif) receptor 5 (CCR5) gene are described. In addition, various genotypes, haplotypes and haplotype pairs for the CCR5 gene that exist in the population are described. Compositions and methods for haplotyping and/or genotyping the CCR5 gene in an individual are also disclosed. Polynucleotides containing one or more of the CCR5 polymorphisms disclosed herein are also described.

WO 01/77125 A2

HAPLOTYPES OF THE CCR5 GENE

RELATED APPLICATIONS

This application claims the benefit of U.S. Provisional Application Serial No. 60/194,361 filed April 5, 2000.

FIELD OF THE INVENTION

This invention relates to variation in genes that encode pharmaceutically-important proteins. In particular, this invention provides genetic variants of the human chemokine (C-C motif) receptor 5 (CCR5) gene and methods for identifying which variant(s) of this gene is/are possessed by an individual.

BACKGROUND OF THE INVENTION

Current methods for identifying pharmaceuticals to treat disease often start by identifying, cloning, and expressing an important target protein related to the disease. A determination of whether an agonist or antagonist is needed to produce an effect that may benefit a patient with the disease is then made. Then, vast numbers of compounds are screened against the target protein to find new potential drugs. The desired outcome of this process is a lead compound that is specific for the target, thereby reducing the incidence of the undesired side effects usually caused by activity at non-intended targets. The lead compound identified in this screening process then undergoes further *in vitro* and *in vivo* testing to determine its absorption, disposition, metabolism and toxicological profiles. Typically, this testing involves use of cell lines and animal models with limited, if any, genetic diversity.

What this approach fails to consider, however, is that natural genetic variability exists between individuals in any and every population with respect to pharmaceutically-important proteins, including the protein targets of candidate drugs, the enzymes that metabolize these drugs and the proteins whose activity is modulated by such drug targets. Subtle alteration(s) in the primary nucleotide sequence of a gene encoding a pharmaceutically-important protein may be manifested as significant variation in expression, structure and/or function of the protein. Such alterations may explain the relatively high degree of uncertainty inherent in the treatment of individuals with a drug whose design is based upon a single representative example of the target or enzyme(s) involved in metabolizing the drug. For example, it is well-established that some drugs frequently have lower efficacy in some individuals than others, which means such individuals and their physicians must weigh the possible benefit of a larger dosage against a greater risk of side effects. Also, there is significant variation in how well people metabolize drugs and other exogenous chemicals, resulting in substantial interindividual variation in the toxicity and/or efficacy of such exogenous substances (Evans et al., 1999, *Science* 286:487-491). This variability in efficacy or toxicity of a drug in genetically-diverse patients makes many drugs ineffective or even dangerous in certain groups of the population, leading to the failure of such drugs in clinical trials or their early withdrawal from the market even though they could be highly beneficial for

other groups in the population. This problem significantly increases the time and cost of drug discovery and development, which is a matter of great public concern.

It is well-recognized by pharmaceutical scientists that considering the impact of the genetic variability of pharmaceutically-important proteins in the early phases of drug discovery and development is likely to reduce the failure rate of candidate and approved drugs (Marshall A 1997 *Nature Biotech* 15:1249-52; Kleyn PW et al. 1998 *Science* 281: 1820-21; Kola I 1999 *Curr Opin Biotech* 10:589-92; Hill AVS et al. 1999 in *Evolution in Health and Disease* Stearns SS (Ed.) Oxford University Press, New York, pp 62-76; Meyer U.A. 1999 in *Evolution in Health and Disease* Stearns SS (Ed.) Oxford University Press, New York, pp 41-49; Kalow W et al. 1999 *Clin. Pharm. Therap.* 66:445-7; Marshall, E 1999 *Science* 284:406-7; Judson R et al. 2000 *Pharmacogenomics* 1:1-12; Roses AD 2000 *Nature* 405:857-65). However, in practice this has been difficult to do, in large part because of the time and cost required for discovering the amount of genetic variation that exists in the population (Chakravarti A 1998 *Nature Genet* 19:216-7; Wang DG et al 1998 *Science* 280:1077-82; Chakravarti A 1999 *Nat Genet* 21:56-60 (suppl); Stephens JC 1999 *Mol. Diagnosis* 4:309-317; Kwok PY and Gu S 1999 *Mol. Med. Today* 5:538-43; Davidson S 2000 *Nature Biotech* 18:1134-5).

The standard for measuring genetic variation among individuals is the haplotype, which is the ordered combination of polymorphisms in the sequence of each form of a gene that exists in the population. Because haplotypes represent the variation across each form of a gene, they provide a more accurate and reliable measurement of genetic variation than individual polymorphisms. For example, while specific variations in gene sequences have been associated with a particular phenotype such as disease susceptibility (Roses AD *supra*; Ulbrecht M et al. 2000 *Am J Respir Crit Care Med* 161: 469-74) and drug response (Wolfe CR et al. 2000 *BMJ* 320:987-90; Dahl BS 1997 *Acta Psychiatr Scand* 96 (Suppl 391): 14-21), in many other cases an individual polymorphism may be found in a variety of genomic backgrounds, i.e., different haplotypes, and therefore shows no definitive coupling between the polymorphism and the causative site for the phenotype (Clark AG et al. 1998 *Am J Hum Genet* 63:595-612; Ulbrecht M et al. 2000 *supra*; Drysdale et al. 2000 *PNAS* 97:10483-10488). Thus, there is an unmet need in the pharmaceutical industry for information on what haplotypes exist in the population for pharmaceutically-important genes. Such haplotype information would be useful in improving the efficiency and output of several steps in the drug discovery and development process, including target validation, identifying lead compounds, and early phase clinical trials (Marshall et al., *supra*).

One pharmaceutically-important gene for the treatment of HIV infection and AIDS is the chemokine (C-C motif) receptor 5 (CCR5) gene or its encoded product. CCR5 is part of a family of trans-membrane proteins that transduce signals via heterotrimeric G-proteins (Horuk, Trends Pharmacol. Sci. 1994; 15:159-165). Chemokines are a family of structurally similar proteins that help elicit a proinflammatory response by binding to receptors on leukocytes. The CXC chemokines, or α chemokines, act on neutrophils, while the CC chemokines act on monocytes, lymphocytes, basophils and eosinophils (Choe et al., Cell 1996; 85:1135-1148). CCR5 was identified as a coreceptor for the

human immunodeficiency virus type 1 (HIV-1), and the presence of this protein renders nonpermissive CD4+ cells susceptible to HIV-1 strains (Deng et al., *Nature* 1996; 381:661-666; Dragic et al., *Nature* 1996; 381:667-673). CCR5 has also been shown to enhance HIV-1 infection in HeLa cells (Choe et al., *Cell* 1996; 85:1135-1148). Although HIV binding and internalization can be mediated by CD4 acting together with several members of the chemokine receptor family, CCR5 appears to be the critical receptor used by HIV glycoproteins during the early stages of infection (OMIM 601373,). Mutations that alter the CCR5 gene product have been associated with resistance to HIV-1. In vitro fusion assays have shown that a deletion resulting in truncated CCR5 do not allow fusion of CD4+ cells with viral propagating cells expressing env protein. The truncated protein was not detected at the surface of cells that normally have the receptor (Samson et al., *Nature* 1996; 382:722-725; Liu et al., *Cell* 1996; 86:367-377). Zimmerman et al. (*Mol. Med.* 1997; 3:23-36) found the CCR5 deletion mutation, which they designated CCR5-2, in random blood donors from North America, Asia, and Africa, and observed that homozygosity for the inactive allele was an HIV-1 resistance factor, while heterozygotes demonstrate a slower rate of disease progression. These data provide a molecular basis for observed differences between individuals in both the susceptibility to HIV-1 infection and progression to AIDS.

The chemokine (C-C motif) receptor 5 gene is located on chromosome 3p21 and contains 2 exons that encode a 352 amino acid protein. A reference sequence for the CCR5 gene is shown in Figure 1 (GenBank Accession No. U95626.1; SEQ ID NO:1). Reference sequences for the coding sequence (GenBank Accession No. NM_000579.1) and protein are shown in Figures 2 (SEQ ID NO:2) and 3 (SEQ ID NO:3), respectively.

Several polymorphisms have been reported in the CCR5 gene. Polymorphisms of guanine or adenine at a position corresponding to nucleotide 59029 (McDermott et al., *Lancet* 1998; 352:866-870); thymine or cytosine at a position corresponding to nucleotide 59353 (SNP ID:rs#1799988); cytosine or thymine at a position corresponding to nucleotide 59356 (Kostrikis et al., *J. Virol.* 1999; 73:10264-10271); guanine or adenine at a position corresponding to nucleotide 59402 (SNP ID:rs#1800023); cytosine or thymine at a position corresponding to nucleotide 59653 (HGBASE:SNP 000002673); thymine or adenine at a position corresponding to nucleotide 61646 (SNP ID:rs#1799863); thymine or cytosine at a position corresponding to nucleotide 61707 (SNP ID:rs#1800941); and guanine or adenine at a position corresponding to nucleotide 62150 (SNP ID:rs#1800452) in Figure 1. Some of these single nucleotide polymorphisms have been associated with phenotypes. The SNP corresponding to nucleotide 59029 in Figure 1 results in a slowed progression to AIDS following seroconversion (McDermott et al., *Lancet* 1998; 352:866-870). The SNPs corresponding to nucleotide 59356 and 59402 in Figure 1 are responsible for increasing the rate of parental HIV-1 transmission (Kostrikis et al., *J. Virol.* 1999; 73:10264-10271) (Kostrikis et al., *Nat. Med.* 1998; 4:350-353).

Because of the potential for variation in the CCR5 gene to affect the expression and function of the encoded protein, it would be useful to know whether additional polymorphisms exist in the

CCR5 gene, as well as how such polymorphisms are combined in different copies of the gene. Such information could be applied for studying the biological function of CCR5 as well as in identifying drugs targeting this protein for the treatment of disorders related to its abnormal expression or function.

SUMMARY OF THE INVENTION

Accordingly, the inventors herein have discovered 4 novel polymorphic sites in the CCR5 gene. These polymorphic sites (PS) correspond to the following nucleotide positions in the indicated GenBank Accession Number: 59802 (PS6), 61370 (PS7), 61803 (PS10) and 62028 (PS11) in U95626.1. The polymorphisms at these sites are adenine or cytosine at PS6, guanine or thymine at PS7, cytosine or thymine at PS10 and thymine or guanine at PS11. In addition, the inventors have determined the identity of the alleles at these sites, as well as at the previously identified sites at nucleotide positions 59029 (PS1), 59353 (PS2), 59356 (PS3), 59402 (PS4), 59653 (PS5), 61646 (PS8), 61707 (PS9) and 62150 (PS12) in U95626.1, in a human reference population of 79 unrelated individuals self-identified as belonging to one of four major population groups: African descent, Asian, Caucasian and Hispanic/Latino. From this information, the inventors deduced a set of haplotypes and haplotype pairs for PS1-PS12 in the CCR5 gene, which are shown below in Tables 5 and 4, respectively. Each of these CCR5 haplotypes defines a naturally-occurring isoform (also referred to herein as an "isogene") of the CCR5 gene that exists in the human population.

Thus, in one embodiment, the invention provides a method, composition and kit for genotyping the CCR5 gene in an individual. The genotyping method comprises identifying the nucleotide pair that is present at one or more polymorphic sites selected from the group consisting of PS6, PS7, PS10 and PS11 in both copies of the CCR5 gene from the individual. A genotyping composition of the invention comprises an oligonucleotide probe or primer which is designed to specifically hybridize to a target region containing, or adjacent to, one of these novel CCR5 polymorphic sites. A genotyping kit of the invention comprises a set of oligonucleotides designed to genotype each of these novel CCR5 polymorphic sites. In a preferred embodiment, the genotyping kit comprises a set of oligonucleotides designed to genotype each of PS1-PS12. The genotyping method, composition, and kit are useful in determining whether an individual has one of the haplotypes in Table 5 below or has one of the haplotype pairs in Table 4 below.

The invention also provides a method for haplotyping the CCR5 gene in an individual. In one embodiment, the haplotyping method comprises determining, for one copy of the CCR5 gene, the identity of the nucleotide at one or more polymorphic sites selected from the group consisting of PS6, PS7, PS10 and PS11. In another embodiment, the haplotyping method comprises determining whether one copy of the individual's CCR5 gene is defined by one of the CCR5 haplotypes shown in Table 5, below, or a sub-haplotype thereof. In a preferred embodiment, the haplotyping method comprises determining whether both copies of the individual's CCR5 gene are defined by one of the CCR5 haplotype pairs shown in Table 4 below, or a sub-haplotype pair thereof. The method for establishing

the CCR5 haplotype or haplotype pair of an individual is useful for improving the efficiency and reliability of several steps in the discovery and development of drugs for treating diseases associated with CCR5 activity, e.g., HIV infection and AIDS.

For example, the haplotyping method can be used by the pharmaceutical research scientist to validate CCR5 as a candidate target for treating a specific condition or disease predicted to be associated with CCR5 activity. Determining for a particular population the frequency of one or more of the individual CCR5 haplotypes or haplotype pairs described herein will facilitate a decision on whether to pursue CCR5 as a target for treating the specific disease of interest. In particular, if variable CCR5 activity is associated with the disease, then one or more CCR5 haplotypes or haplotype pairs will be found at a higher frequency in disease cohorts than in appropriately genetically matched controls. Conversely, if each of the observed CCR5 haplotypes are of similar frequencies in the disease and control groups, then it may be inferred that variable CCR5 activity has little, if any, involvement with that disease. In either case, the pharmaceutical research scientist can, without *a priori* knowledge as to the phenotypic effect of any CCR5 haplotype or haplotype pair, apply the information derived from detecting CCR5 haplotypes in an individual to decide whether modulating CCR5 activity would be useful in treating the disease.

The claimed invention is also useful in screening for compounds targeting CCR5 to treat a specific condition or disease predicted to be associated with CCR5 activity. For example, detecting which of the CCR5 haplotypes or haplotype pairs disclosed herein are present in individual members of a population with the specific disease of interest enables the pharmaceutical scientist to screen for a compound(s) that displays the highest desired agonist or antagonist activity for each of the most frequent CCR5 isoforms present in the disease population. Thus, without requiring any *a priori* knowledge of the phenotypic effect of any particular CCR5 haplotype or haplotype pair, the claimed haplotyping method provides the scientist with a tool to identify lead compounds that are more likely to show efficacy in clinical trials.

The method for haplotyping the CCR5 gene in an individual is also useful in the design of clinical trials of candidate drugs for treating a specific condition or disease predicted to be associated with CCR5 activity. For example, instead of randomly assigning patients with the disease of interest to the treatment or control group as is typically done now, determining which of the CCR5 haplotype(s) disclosed herein are present in individual patients enables the pharmaceutical scientist to distribute CCR5 haplotypes and/or haplotype pairs evenly to treatment and control groups, thereby reducing the potential for bias in the results that could be introduced by a larger frequency of a CCR5 haplotype or haplotype pair that had a previously unknown association with response to the drug being studied in the trial. Thus, by practicing the claimed invention, the scientist can more confidently rely on the information learned from the trial, without first determining the phenotypic effect of any CCR5 haplotype or haplotype pair.

In another embodiment, the invention provides a method for identifying an association between a trait and a CCR5 genotype, haplotype, or haplotype pair for one or more of the novel

polymorphic sites described herein. The method comprises comparing the frequency of the CCR5 genotype, haplotype, or haplotype pair in a population exhibiting the trait with the frequency of the CCR5 genotype, haplotype, or haplotype pair in a reference population. A higher frequency of the CCR5 genotype, haplotype, or haplotype pair in the trait population than in the reference population indicates the trait is associated with the CCR5 genotype, haplotype, or haplotype pair. In preferred embodiments, the trait is susceptibility to a disease, severity of a disease, the staging of a disease or response to a drug. In a particularly preferred embodiment, the CCR5 haplotype is selected from the haplotypes shown in Table 5, or a sub-haplotype thereof. Such methods have applicability in developing diagnostic tests and therapeutic treatments for HIV infection and AIDS.

In yet another embodiment, the invention provides an isolated polynucleotide comprising a nucleotide sequence which is a polymorphic variant of a reference sequence for the CCR5 gene or a fragment thereof. The reference sequence comprises SEQ ID NO:1 and the polymorphic variant comprises at least one polymorphism selected from the group consisting of cytosine at PS6, thymine at PS7, thymine at PS10 and guanine at PS11. In a preferred embodiment, the polymorphic variant comprises one or more additional polymorphisms selected from the group consisting of adenine at PS1, cytosine at PS2, thymine at PS3, adenine at PS4, thymine at PS5, adenine at PS8, cytosine at PS9 and adenine at PS12.

A particularly preferred polymorphic variant is an isogene of the CCR5 gene. A CCR5 isogene of the invention comprises guanine or adenine at PS1, thymine or cytosine at PS2, cytosine or thymine at PS3, guanine or adenine at PS4, cytosine or thymine at PS5, adenine or cytosine at PS6, guanine or thymine at PS7, thymine or adenine at PS8, thymine or cytosine at PS9, cytosine or thymine at PS10, thymine or guanine at PS11 and guanine or adenine at PS12. The invention also provides a collection of CCR5 isogenes, referred to herein as a CCR5 genome anthology.

In another embodiment, the invention provides a polynucleotide comprising a polymorphic variant of a reference sequence for a CCR5 cDNA or a fragment thereof. The reference sequence comprises SEQ ID NO:2 (Fig.2) and the polymorphic cDNA comprises at least one polymorphism selected from the group consisting of thymine at a position corresponding to nucleotide 321 and guanine at a position corresponding to nucleotide 546. In a preferred embodiment, the polymorphic variant comprises one or more additional polymorphisms selected from the group consisting of adenine at a position corresponding to nucleotide 164, cytosine at a position corresponding to nucleotide 225 and adenine at a position corresponding to nucleotide 668. A particularly preferred polymorphic cDNA variant comprises the coding sequence of a CCR5 isogene defined by haplotypes 2-11.

Polynucleotides complementary to these CCR5 genomic and cDNA variants are also provided by the invention. It is believed that polymorphic variants of the CCR5-gene will be useful in studying the expression and function of CCR5, and in expressing CCR5 protein for use in screening for candidate drugs to treat diseases related to CCR5 activity.

In other embodiments, the invention provides a recombinant expression vector comprising one of the polymorphic genomic variants operably linked to expression regulatory elements as well as a

recombinant host cell transformed or transfected with the expression vector. The recombinant vector and host cell may be used to express CCR5 for protein structure analysis and drug binding studies.

In yet another embodiment, the invention provides a polypeptide comprising a polymorphic variant of a reference amino acid sequence for the CCR5 protein. The reference amino acid sequence comprises SEQ ID NO:3 (Fig.3) and the polymorphic variant comprises leucine at a position corresponding to amino acid position 182. In some embodiments, the polymorphic variant also comprises at least one variant amino acid selected from the group consisting of glutamine at a position corresponding to amino acid position 55 and glutamine at a position corresponding to amino acid position 223. A polymorphic variant of CCR5 is useful in studying the effect of the variation on the biological activity of CCR5 as well as on the binding affinity of candidate drugs targeting CCR5 for the treatment of HIV infection and AIDS.

The present invention also provides antibodies that recognize and bind to the above polymorphic CCR5 protein variant. Such antibodies can be utilized in a variety of diagnostic and prognostic formats and therapeutic methods.

The present invention also provides nonhuman transgenic animals comprising one of the CCR5 polymorphic genomic variants described herein and methods for producing such animals. The transgenic animals are useful for studying expression of the CCR5 isogenes *in vivo*, for *in vivo* screening and testing of drugs targeted against CCR5 protein, and for testing the efficacy of therapeutic agents and compounds for HIV infection and AIDS in a biological system.

The present invention also provides a computer system for storing and displaying polymorphism data determined for the CCR5 gene. The computer system comprises a computer processing unit; a display; and a database containing the polymorphism data. The polymorphism data includes the polymorphisms, the genotypes and the haplotypes identified for the CCR5 gene in a reference population. In a preferred embodiment, the computer system is capable of producing a display showing CCR5 haplotypes organized according to their evolutionary relationships.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 illustrates a reference sequence for the CCR5 gene (Genbank Accession Number U95626.1; contiguous lines; SEQ ID NO:1), with the start and stop positions of each region of coding sequence indicated with a bracket ([or]) and the numerical position below the sequence and the polymorphic site(s) and polymorphism(s) identified by Applicants in a reference population indicated by the variant nucleotide positioned below the polymorphic site in the sequence. SEQ ID NO: 24 is equivalent to Figure 1, with the two alternative allelic variants of each polymorphic site indicated by the appropriate nucleotide symbol (R = G or A, Y = T or C, M = A or C, K = G or T, S = G or C, and W = A or T; WIPO standard ST.25).

Figure 2 illustrates a reference sequence for the CCR5 coding sequence (contiguous lines; SEQ ID NO:2) with the polymorphic site(s) and polymorphism(s) identified by Applicants in a reference population indicated by the variant nucleotide positioned below the polymorphic site in the

sequence.

Figure 3 illustrates a reference sequence for the CCR5 protein (contiguous lines; SEQ ID NO:3), with the variant amino acid(s) caused by the polymorphism(s) of Figure 2 positioned below the polymorphic site in the sequence.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

The present invention is based on the discovery of novel variants of the CCR5 gene. As described in more detail below, the inventors herein discovered 11 isogenes of the CCR5 gene by characterizing the CCR5 gene found in genomic DNAs isolated from an Index Repository that contains immortalized cell lines from one chimpanzee and 93 human individuals. The human individuals included a reference population of 79 unrelated individuals self-identified as belonging to one of four major population groups: Caucasian (CA)(22 individuals), African descent (AF)(20 individuals), Asian (AS)(20 individuals), or Hispanic/Latino (17 individuals)(HL). To the extent possible, the members of this reference population were organized into population subgroups by the self-identified ethnogeographic origin of their four grandparents as shown in Table 1 below.

Table 1. Population Groups in the Index Repository

Population Group	Population Subgroup	No. of Individuals
African descent		20
	Sierra Leone	1
Asian		20
	Burma	1
	China	3
	Japan	6
	Korea	1
	Philippines	5
	Vietnam	4
Caucasian		22
	British Isles	3
	British Isles/Central	4
	British Isles/Eastern	1
	Central/Eastern	1
	Eastern	3
	Central/Mediterranean	1
	Mediterranean	2
	Scandinavian	2
Hispanic/Latino		17
	Caribbean	7
	Caribbean (Spanish Descent)	2
	Central American (Spanish Descent)	1
	Mexican American	4
	South American (Spanish Descent)	3

In addition, the Index Repository contains three unrelated indigenous American Indians (AM)(one from each of North, Central and South America), one three-generation Caucasian family (from the CEPH Utah cohort) and one two-generation African-American family.

The CCR5 isogenes present in the human reference population are defined by haplotypes for 12 polymorphic sites in the CCR5 gene, 4 of which are believed to be novel. The CCR5 polymorphic sites identified by the inventors are referred to as PS1-PS12 to designate the order in which they are located in the gene (see Table 3 below), with the novel polymorphic sites referred to as PS6, PS7, PS10 and PS11. Using the genotypes identified in the Index Repository for PS1-PS12 and the methodology described in the Examples below, the inventors herein also determined the pair of haplotypes for the CCR5 gene present in individual human members of this repository. The human genotypes and haplotypes found in the repository for the CCR5 gene include those shown in Tables 4 and 5, respectively. The polymorphism and haplotype data disclosed herein are useful for validating whether CCR5 is a suitable target for drugs to treat HIV infection and AIDS, screening for such drugs and reducing bias in clinical trials of such drugs.

In the context of this disclosure, the following terms shall be defined as follows unless otherwise indicated:

Allele - A particular form of a genetic locus, distinguished from other forms by its particular nucleotide sequence.

Candidate Gene - A gene which is hypothesized to be responsible for a disease, condition, or the response to a treatment, or to be correlated with one of these.

Gene - A segment of DNA that contains all the information for the regulated biosynthesis of an RNA product, including promoters, exons, introns, and other untranslated regions that control expression.

Genotype - An unphased 5' to 3' sequence of nucleotide pair(s) found at one or more polymorphic sites in a locus on a pair of homologous chromosomes in an individual. As used herein, genotype includes a full-genotype and/or a sub-genotype as described below.

Full-genotype - The unphased 5' to 3' sequence of nucleotide pairs found at all known polymorphic sites in a locus on a pair of homologous chromosomes in a single individual.

Sub-genotype - The unphased 5' to 3' sequence of nucleotides seen at a subset of the known polymorphic sites in a locus on a pair of homologous chromosomes in a single individual.

Genotyping - A process for determining a genotype of an individual.

Haplotype - A 5' to 3' sequence of nucleotides found at one or more polymorphic sites in a locus on a single chromosome from a single individual. As used herein, haplotype includes a full-haplotype and/or a sub-haplotype as described below.

Full-haplotype - The 5' to 3' sequence of nucleotides found at all known polymorphic sites in a locus on a single chromosome from a single individual.

Sub-haplotype - The 5' to 3' sequence of nucleotides seen at a subset of the known polymorphic sites in a locus on a single chromosome from a single individual.

Haplotype pair - The two haplotypes found for a locus in a single individual.

Haplotyping - A process for determining one or more haplotypes in an individual and includes use of family pedigrees, molecular techniques and/or statistical inference.

Haplotype data - Information concerning one or more of the following for a specific gene: a listing of the haplotype pairs in each individual in a population; a listing of the different haplotypes in a population; frequency of each haplotype in that or other populations, and any known associations between one or more haplotypes and a trait.

Isoform - A particular form of a gene, mRNA, cDNA or the protein encoded thereby, distinguished from other forms by its particular sequence and/or structure.

Isogene - One of the isoforms of a gene found in a population. An isogene contains all of the polymorphisms present in the particular isoform of the gene.

Isolated - As applied to a biological molecule such as RNA, DNA, oligonucleotide, or protein, isolated means the molecule is substantially free of other biological molecules such as nucleic acids, proteins, lipids, carbohydrates, or other material such as cellular debris and growth media. Generally, the term "isolated" is not intended to refer to a complete absence of such material or to absence of water, buffers, or salts, unless they are present in amounts that substantially interfere with the methods of the present invention.

Locus - A location on a chromosome or DNA molecule corresponding to a gene or a physical or phenotypic feature.

Naturally-occurring - A term used to designate that the object it is applied to, e.g., naturally-occurring polynucleotide or polypeptide, can be isolated from a source in nature and which has not been intentionally modified by man.

Nucleotide pair - The nucleotides found at a polymorphic site on the two copies of a chromosome from an individual.

Phased - As applied to a sequence of nucleotide pairs for two or more polymorphic sites in a locus, phased means the combination of nucleotides present at those polymorphic sites on a single copy of the locus is known.

Polymorphic site (PS) - A position within a locus at which at least two alternative sequences are found in a population, the most frequent of which has a frequency of no more than 99%.

Polymorphic variant - A gene, mRNA, cDNA, polypeptide or peptide whose nucleotide or amino acid sequence varies from a reference sequence due to the presence of a polymorphism in the gene.

Polymorphism - The sequence variation observed in an individual at a polymorphic site. Polymorphisms include nucleotide substitutions, insertions, deletions and microsatellites and may, but need not, result in detectable differences in gene expression or protein function.

Polymorphism data - Information concerning one or more of the following for a specific gene: location of polymorphic sites; sequence variation at those sites; frequency of polymorphisms in one or more populations; the different genotypes and/or haplotypes determined for the gene; frequency of one or more of these genotypes and/or haplotypes in one or more populations; any known association(s) between a trait and a genotype or a haplotype for the gene.

Polymorphism Database - A collection of polymorphism data arranged in a systematic or

methodical way and capable of being individually accessed by electronic or other means.

Polynucleotide – A nucleic acid molecule comprised of single-stranded RNA or DNA or comprised of complementary, double-stranded DNA.

Population Group – A group of individuals sharing a common ethnogeographic origin.

Reference Population – A group of subjects or individuals who are predicted to be representative of the genetic variation found in the general population. Typically, the reference population represents the genetic variation in the population at a certainty level of at least 85%, preferably at least 90%, more preferably at least 95% and even more preferably at least 99%.

Single Nucleotide Polymorphism (SNP) – Typically, the specific pair of nucleotides observed at a single polymorphic site. In rare cases, three or four nucleotides may be found.

Subject – A human individual whose genotypes or haplotypes or response to treatment or disease state are to be determined.

Treatment – A stimulus administered internally or externally to a subject.

Unphased – As applied to a sequence of nucleotide pairs for two or more polymorphic sites in a locus, unphased means the combination of nucleotides present at those polymorphic sites on a single copy of the locus is not known.

As discussed above, information on the identity of genotypes and haplotypes for the CCR5 gene of any particular individual as well as the frequency of such genotypes and haplotypes in any particular population of individuals is expected to be useful for a variety of drug discovery and development applications. Thus, the invention also provides compositions and methods for detecting the novel CCR5 polymorphisms and haplotypes identified herein.

The compositions comprise at least one CCR5 genotyping oligonucleotide. In one embodiment, a CCR5 genotyping oligonucleotide is a probe or primer capable of hybridizing to a target region that is located close to, or that contains, one of the novel polymorphic sites described herein. As used herein, the term "oligonucleotide" refers to a polynucleotide molecule having less than about 100 nucleotides. A preferred oligonucleotide of the invention is 10 to 35 nucleotides long. More preferably, the oligonucleotide is between 15 and 30, and most preferably, between 20 and 25 nucleotides in length. The exact length of the oligonucleotide will depend on many factors that are routinely considered and practiced by the skilled artisan. The oligonucleotide may be comprised of any phosphorylation state of ribonucleotides, deoxyribonucleotides, and acyclic nucleotide derivatives, and other functionally equivalent derivatives. Alternatively, oligonucleotides may have a phosphate-free backbone, which may be comprised of linkages such as carboxymethyl, acetamidate, carbamate, polyamide (peptide nucleic acid (PNA)) and the like (Varma, R. in *Molecular Biology and Biotechnology*, A Comprehensive Desk Reference, Ed. R. Meyers, VCH Publishers, Inc. (1995), pages 617-620). Oligonucleotides of the invention may be prepared by chemical synthesis using any suitable methodology known in the art, or may be derived from a biological sample, for example, by restriction digestion. The oligonucleotides may be labeled, according to any technique known in the art, including use of radiolabels, fluorescent labels, enzymatic labels, proteins, haptens, antibodies,

sequence tags and the like.

Genotyping oligonucleotides of the invention must be capable of specifically hybridizing to a target region of a CCR5 polynucleotide, i.e., a CCR5 isogene. As used herein, specific hybridization means the oligonucleotide forms an anti-parallel double-stranded structure with the target region under certain hybridizing conditions, while failing to form such a structure when incubated with a non-target region or a non-CCR5 polynucleotide under the same hybridizing conditions. Preferably, the oligonucleotide specifically hybridizes to the target region under conventional high stringency conditions. The skilled artisan can readily design and test oligonucleotide probes and primers suitable for detecting polymorphisms in the CCR5 gene using the polymorphism information provided herein in conjunction with the known sequence information for the CCR5 gene and routine techniques.

A nucleic acid molecule such as an oligonucleotide or polynucleotide is said to be a "perfect" or "complete" complement of another nucleic acid molecule if every nucleotide of one of the molecules is complementary to the nucleotide at the corresponding position of the other molecule. A nucleic acid molecule is "substantially complementary" to another molecule if it hybridizes to that molecule with sufficient stability to remain in a duplex form under conventional low-stringency conditions. Conventional hybridization conditions are described, for example, by Sambrook J. et al., in *Molecular Cloning, A Laboratory Manual*, 2nd Edition, Cold Spring Harbor Press, Cold Spring Harbor, NY (1989) and by Haymes, B.D. et al. in *Nucleic Acid Hybridization, A Practical Approach*, IRL Press, Washington, D.C. (1985). While perfectly complementary oligonucleotides are preferred for detecting polymorphisms, departures from complete complementarity are contemplated where such departures do not prevent the molecule from specifically hybridizing to the target region. For example, an oligonucleotide primer may have a non-complementary fragment at its 5' end, with the remainder of the primer being complementary to the target region. Alternatively, non-complementary nucleotides may be interspersed into the oligonucleotide probe or primer as long as the resulting probe or primer is still capable of specifically hybridizing to the target region.

Preferred genotyping oligonucleotides of the invention are allele-specific oligonucleotides. As used herein, the term allele-specific oligonucleotide (ASO) means an oligonucleotide that is able, under sufficiently stringent conditions, to hybridize specifically to one allele of a gene, or other locus, at a target region containing a polymorphic site while not hybridizing to the corresponding region in another allele(s). As understood by the skilled artisan, allele-specificity will depend upon a variety of readily optimized stringency conditions, including salt and formamide concentrations, as well as temperatures for both the hybridization and washing steps. Examples of hybridization and washing conditions typically used for ASO probes are found in Kogan et al., "Genetic Prediction of Hemophilia A" in *PCR Protocols, A Guide to Methods and Applications*, Academic Press, 1990 and Ruaño et al., 87 *Proc. Natl. Acad. Sci. USA* 6296-6300, 1990. Typically, an ASO will be perfectly complementary to one allele while containing a single mismatch for another allele.

Allele-specific oligonucleotides of the invention include ASO probes and ASO primers. ASO probes which usually provide good discrimination between different alleles are those in which a central

position of the oligonucleotide probe aligns with the polymorphic site in the target region (e.g., approximately the 7th or 8th position in a 15mer, the 8th or 9th position in a 16mer, and the 10th or 11th position in a 20mer). An ASO primer of the invention has a 3' terminal nucleotide, or preferably a 3' penultimate nucleotide, that is complementary to only one nucleotide of a particular SNP, thereby acting as a primer for polymerase-mediated extension only if the allele containing that nucleotide is present. ASO probes and primers hybridizing to either the coding or noncoding strand are contemplated by the invention.

ASO probes and primers listed below use the appropriate nucleotide symbol (R= G or A, Y= T or C, M= A or C, K= G or T, S= G or C, and W= A or T; WIPO standard ST.25) at the position of the polymorphic site to represent the two alternative allelic variants observed at that polymorphic site.

A preferred ASO probe for detecting CCR5 gene polymorphisms comprises a nucleotide sequence, listed 5' to 3', selected from the group consisting of:

Accession No.: U95626.1

GAGCAATMGATTTTT (SEQ ID NO:4) and its complement,
CTATGTAKGCAATTA (SEQ ID NO:5) and its complement,
CAGGGCTYTATTTTA (SEQ ID NO:6) and its complement, and
CTCATTTKCCATACA (SEQ ID NO:7) and its complement.

A preferred ASO primer for detecting CCR5 gene polymorphisms comprises a nucleotide sequence, listed 5' to 3', selected from the group consisting of:

Accession No.:U95626.1.

TAAAGGGAGCAATMG (SEQ ID NO:8); GTTATTAAAATACKA (SEQ ID NO:9);
AGACATCTATGTAKG (SEQ ID NO:10); GGTTTTTAATTGCMT (SEQ ID NO:11);
TCTTGACAGGGCTYT (SEQ ID NO:12); AGCCTATAAAATARA (SEQ ID NO:13);
GCAGCTCTCATTTKC (SEQ ID NO:14); and ACTGACTGTATGGMA (SEQ ID NO:15).

Other genotyping oligonucleotides of the invention hybridize to a target region located one to several nucleotides downstream of one of the novel polymorphic sites identified herein. Such oligonucleotides are useful in polymerase-mediated primer extension methods for detecting one of the novel polymorphisms described herein and therefore such genotyping oligonucleotides are referred to herein as "primer-extension oligonucleotides". In a preferred embodiment, the 3'-terminus of a primer-extension oligonucleotide is a deoxynucleotide complementary to the nucleotide located immediately adjacent to the polymorphic site.

A particularly preferred oligonucleotide primer for detecting CCR5 gene polymorphisms by primer extension terminates in a nucleotide sequence, listed 5' to 3', selected from the group consisting of:

Accession No.:U95626.1

AGGGAGCAAT (SEQ ID NO:16); ATTAAAATAC (SEQ ID NO:17);

CATCTATGTA (SEQ ID NO:18); TTTTAATTGC (SEQ ID NO:19);
TGACAGGGCT (SEQ ID NO:20); CTATAAAATA (SEQ ID NO:21);
GCTCTCATTT (SEQ ID NO:22); and GACTGTATGG (SEQ ID NO:23).

In some embodiments, a composition contains two or more differently labeled genotyping oligonucleotides for simultaneously probing the identity of nucleotides at two or more polymorphic sites. It is also contemplated that primer compositions may contain two or more sets of allele-specific primer pairs to allow simultaneous targeting and amplification of two or more regions containing a polymorphic site.

CCR5 genotyping oligonucleotides of the invention may also be immobilized on or synthesized on a solid surface such as a microchip, bead, or glass slide (see, e.g., WO 98/20020 and WO 98/20019). Such immobilized genotyping oligonucleotides may be used in a variety of polymorphism detection assays, including but not limited to probe hybridization and polymerase extension assays. Immobilized CCR5 genotyping oligonucleotides of the invention may comprise an ordered array of oligonucleotides designed to rapidly screen a DNA sample for polymorphisms in multiple genes at the same time.

In another embodiment, the invention provides a kit comprising at least two genotyping oligonucleotides packaged in separate containers. The kit may also contain other components such as hybridization buffer (where the oligonucleotides are to be used as a probe) packaged in a separate container. Alternatively, where the oligonucleotides are to be used to amplify a target region, the kit may contain, packaged in separate containers, a polymerase and a reaction buffer optimized for primer extension mediated by the polymerase, such as PCR.

The above described oligonucleotide compositions and kits are useful in methods for genotyping and/or haplotyping the CCR5 gene in an individual. As used herein, the terms "CCR5 genotype" and "CCR5 haplotype" mean the genotype or haplotype contains the nucleotide pair or nucleotide, respectively, that is present at one or more of the novel polymorphic sites described herein and may optionally also include the nucleotide pair or nucleotide present at one or more additional polymorphic sites in the CCR5 gene. The additional polymorphic sites may be currently known polymorphic sites or sites that are subsequently discovered.

One embodiment of the genotyping method involves isolating from the individual a nucleic acid sample comprising the two copies of the CCR5 gene, or a fragment thereof, that are present in the individual, and determining the identity of the nucleotide pair at one or more polymorphic sites selected from the group consisting of PS6, PS7, PS10 and PS11 in the two copies to assign a CCR5 genotype to the individual. As will be readily understood by the skilled artisan, the two "copies" of a gene in an individual may be the same allele or may be different alleles. In a preferred embodiment of the genotyping method, the identity of the nucleotide pair at one or more of the polymorphic sites selected from the group consisting of PS1, PS2, PS3, PS4, PS5, PS8, PS9 and PS12 is also determined. In a particularly preferred embodiment, the genotyping method comprises determining the identity of

the nucleotide pair at each of PS1-PS12.

Typically, the nucleic acid sample is isolated from a biological sample taken from the individual, such as a blood sample or tissue sample. Suitable tissue samples include whole blood, semen, saliva, tears, urine, fecal material, sweat, buccal, skin and hair. The nucleic acid sample may be comprised of genomic DNA, mRNA, or cDNA and, in the latter two cases, the biological sample must be obtained from a tissue in which the CCR5 gene is expressed. Furthermore it will be understood by the skilled artisan that mRNA or cDNA preparations would not be used to detect polymorphisms located in introns or in 5' and 3' untranslated regions. If a CCR5 gene fragment is isolated, it must contain the polymorphic site(s) to be genotyped.

One embodiment of the haplotyping method comprises isolating from the individual a nucleic acid sample containing only one of the two copies of the CCR5 gene, or a fragment thereof, that is present in the individual and determining in that copy the identity of the nucleotide at one or more polymorphic sites selected from the group consisting of PS6, PS7, PS10 and PS11 in that copy to assign a CCR5 haplotype to the individual. The nucleic acid may be isolated using any method capable of separating the two copies of the CCR5 gene or fragment such as one of the methods described above for preparing CCR5 isogenes, with targeted *in vivo* cloning being the preferred approach. As will be readily appreciated by those skilled in the art, any individual clone will only provide haplotype information on one of the two CCR5 gene copies present in an individual. If haplotype information is desired for the individual's other copy, additional CCR5 clones will need to be examined. Typically, at least five clones should be examined to have more than a 90% probability of haplotyping both copies of the CCR5 gene in an individual. In some embodiments, the haplotyping method also comprises identifying the nucleotide at one or more polymorphic sites selected from the group consisting of PS1, PS2, PS3, PS4, PS5, PS8, PS9 and PS12. In a particularly preferred embodiment, the nucleotide at each of PS1-PS12 is identified.

In another embodiment, the haplotyping method comprises determining whether an individual has one or more of the CCR5 haplotypes shown in Table 5. This can be accomplished by identifying, for one or both copies of the individual's CCR5 gene, the phased sequence of nucleotides present at each of PS1-PS12. The present invention also contemplates that typically only a subset of PS1-PS12 will need to be directly examined to assign to an individual one or more of the haplotypes shown in Table 5. This is because at least one polymorphic site in a gene is frequently in strong linkage disequilibrium with one or more other polymorphic sites in that gene (Drysedale, CM et al. 2000 *PNAS* 97:10483-10488; Rieder MJ et al. 1999 *Nature Genetics* 22:59-62). Two sites are said to be in linkage disequilibrium if the presence of a particular variant at one site enhances the predictability of another variant at the second site (Stephens, JC 1999, *Mol. Diag.* 4:309-317). Techniques for determining whether any two polymorphic sites are in linkage disequilibrium are well-known in the art (Weir B.S. 1996 *Genetic Data Analysis II*, Sinauer Associates, Inc. Publishers, Sunderland, MA).

In a preferred embodiment, a CCR5 haplotype pair is determined for an individual by identifying the phased sequence of nucleotides at one or more polymorphic sites selected from the

group consisting of PS6, PS7, PS10 and PS11 in each copy of the CCR5 gene that is present in the individual. In a particularly preferred embodiment, the haplotyping method comprises identifying the phased sequence of nucleotides at each of PS1-PS12 in each copy of the CCR5 gene. When haplotyping both copies of the gene, the identifying step is preferably performed with each copy of the gene being placed in separate containers. However, it is also envisioned that if the two copies are labeled with different tags, or are otherwise separately distinguishable or identifiable, it could be possible in some cases to perform the method in the same container. For example, if first and second copies of the gene are labeled with different first and second fluorescent dyes, respectively, and an allele-specific oligonucleotide labeled with yet a third different fluorescent dye is used to assay the polymorphic site(s), then detecting a combination of the first and third dyes would identify the polymorphism in the first gene copy while detecting a combination of the second and third dyes would identify the polymorphism in the second gene copy.

In both the genotyping and haplotyping methods, the identity of a nucleotide (or nucleotide pair) at a polymorphic site(s) may be determined by amplifying a target region(s) containing the polymorphic site(s) directly from one or both copies of the CCR5 gene, or a fragment thereof, and the sequence of the amplified region(s) determined by conventional methods. It will be readily appreciated by the skilled artisan that only one nucleotide will be detected at a polymorphic site in individuals who are homozygous at that site, while two different nucleotides will be detected if the individual is heterozygous for that site. The polymorphism may be identified directly, known as positive-type identification, or by inference, referred to as negative-type identification. For example, where a SNP is known to be guanine and cytosine in a reference population, a site may be positively determined to be either guanine or cytosine for an individual homozygous at that site, or both guanine and cytosine, if the individual is heterozygous at that site. Alternatively, the site may be negatively determined to be not guanine (and thus cytosine/cytosine) or not cytosine (and thus guanine/guanine).

The target region(s) may be amplified using any oligonucleotide-directed amplification method, including but not limited to polymerase chain reaction (PCR) (U.S. Patent No. 4,965,188), ligase chain reaction (LCR) (Barany et al., *Proc. Natl. Acad. Sci. USA* 88:189-193, 1991; WO90/01069), and oligonucleotide ligation assay (OLA) (Landegren et al., *Science* 241:1077-1080, 1988).

Other known nucleic acid amplification procedures may be used to amplify the target region including transcription-based amplification systems (U.S. Patent No. 5,130,238; EP 329,822; U.S. Patent No. 5,169,766, WO89/06700) and isothermal methods (Walker et al., *Proc. Natl. Acad. Sci. USA* 89:392-396, 1992).

A polymorphism in the target region may also be assayed before or after amplification using one of several hybridization-based methods known in the art. Typically, allele-specific oligonucleotides are utilized in performing such methods. The allele-specific oligonucleotides may be used as differently labeled probe pairs, with one member of the pair showing a perfect match to one variant of a target sequence and the other member showing a perfect match to a different variant. In

some embodiments, more than one polymorphic site may be detected at once using a set of allele-specific oligonucleotides or oligonucleotide pairs. Preferably, the members of the set have melting temperatures within 5°C, and more preferably within 2°C, of each other when hybridizing to each of the polymorphic sites being detected.

Hybridization of an allele-specific oligonucleotide to a target polynucleotide may be performed with both entities in solution, or such hybridization may be performed when either the oligonucleotide or the target polynucleotide is covalently or noncovalently affixed to a solid support. Attachment may be mediated, for example, by antibody-antigen interactions, poly-L-Lys, streptavidin or avidin-biotin, salt bridges, hydrophobic interactions, chemical linkages, UV cross-linking baking, etc. Allele-specific oligonucleotides may be synthesized directly on the solid support or attached to the solid support subsequent to synthesis. Solid-supports suitable for use in detection methods of the invention include substrates made of silicon, glass, plastic, paper and the like, which may be formed, for example, into wells (as in 96-well plates), slides, sheets, membranes, fibers, chips, dishes, and beads. The solid support may be treated, coated or derivatized to facilitate the immobilization of the allele-specific oligonucleotide or target nucleic acid.

The genotype or haplotype for the CCR5 gene of an individual may also be determined by hybridization of a nucleic acid sample containing one or both copies of the gene, or fragment(s) thereof, to nucleic acid arrays and subarrays such as described in WO 95/11995. The arrays would contain a battery of allele-specific oligonucleotides representing each of the polymorphic sites to be included in the genotype or haplotype.

The identity of polymorphisms may also be determined using a mismatch detection technique, including but not limited to the RNase protection method using riboprobes (Winter et al., *Proc. Natl. Acad. Sci. USA* 82:7575, 1985; Meyers et al., *Science* 230:1242, 1985) and proteins which recognize nucleotide mismatches, such as the E. coli mutS protein (Modrich, *P. Am. Rev. Genet.* 25:229-253, 1991). Alternatively, variant alleles can be identified by single strand conformation polymorphism (SSCP) analysis (Orita et al., *Genomics* 5:874-879, 1989; Humphries et al., in *Molecular Diagnosis of Genetic Diseases*, R. Elles, ed., pp. 321-340, 1996) or denaturing gradient gel electrophoresis (DGGE) (Wartell et al., *Nucl. Acids Res.* 18:2699-2706, 1990; Sheffield et al., *Proc. Natl. Acad. Sci. USA* 86:232-236, 1989).

A polymerase-mediated primer extension method may also be used to identify the polymorphism(s). Several such methods have been described in the patent and scientific literature and include the "Genetic Bit Analysis" method (WO92/15712) and the ligase/polymerase mediated genetic bit analysis (U.S. Patent 5,679,524. Related methods are disclosed in WO91/02087, WO90/09455, WO95/17676, U.S. Patent Nos. 5,302,509, and 5,945,283. Extended primers containing a polymorphism may be detected by mass spectrometry as described in U.S. Patent No. 5,605,798. Another primer extension method is allele-specific PCR (Ruano et al., *Nucl. Acids Res.* 17:8392, 1989; Ruano et al., *Nucl. Acids Res.* 19, 6877-6882, 1991; WO 93/22456; Turki et al., *J. Clin. Invest.* 95:1635-1641, 1995). In addition, multiple polymorphic sites may be investigated by simultaneously

amplifying multiple regions of the nucleic acid using sets of allele-specific primers as described in Wallace et al. (WO89/10414).

In addition, the identity of the allele(s) present at any of the novel polymorphic sites described herein may be indirectly determined by genotyping another polymorphic site that is in linkage disequilibrium with the polymorphic site that is of interest. Polymorphic sites in linkage disequilibrium with the presently disclosed polymorphic sites may be located in regions of the gene or in other genomic regions not examined herein. Genotyping of a polymorphic site in linkage disequilibrium with the novel polymorphic sites described herein may be performed by, but is not limited to, any of the above-mentioned methods for detecting the identity of the allele at a polymorphic site.

In another aspect of the invention, an individual's CCR5 haplotype pair is predicted from its CCR5 genotype using information on haplotype pairs known to exist in a reference population. In its broadest embodiment, the haplotyping prediction method comprises identifying a CCR5 genotype for the individual at two or more CCR5 polymorphic sites described herein, enumerating all possible haplotype pairs which are consistent with the genotype, accessing data containing CCR5 haplotype pairs identified in a reference population, and assigning a haplotype pair to the individual that is consistent with the data. In one embodiment, the reference haplotype pairs include the CCR5 haplotype pairs shown in Table 4.

Generally, the reference population should be composed of randomly-selected individuals representing the major ethnogeographic groups of the world. A preferred reference population for use in the methods of the present invention comprises an approximately equal number of individuals from Caucasian, African American, Asian and Hispanic-Latino population groups with the minimum number of each group being chosen based on how rare a haplotype one wants to be guaranteed to see. For example, if one wants to have a $q\%$ chance of not missing a haplotype that exists in the population at a $p\%$ frequency of occurring in the reference population, the number of individuals (n) who must be sampled is given by $2n = \log(1-q)/\log(1-p)$ where p and q are expressed as fractions. A preferred reference population allows the detection of any haplotype whose frequency is at least 10% with about 99% certainty and comprises about 20 unrelated individuals from each of the four population groups named above. A particularly preferred reference population includes a 3-generation family representing one or more of the four population groups to serve as controls for checking quality of haplotyping procedures.

In a preferred embodiment, the haplotype frequency data for each ethnogeographic group is examined to determine whether it is consistent with Hardy-Weinberg equilibrium. Hardy-Weinberg equilibrium (D.L. Hartl et al., Principles of Population Genomics, Sinauer Associates (Sunderland, MA), 3rd Ed., 1997) postulates that the frequency of finding the haplotype pair H_1 / H_2 is equal to $p_{H-W}(H_1 / H_2) = 2p(H_1)p(H_2)$ if $H_1 \neq H_2$ and $p_{H-W}(H_1 / H_2) = p(H_1)p(H_2)$ if $H_1 = H_2$. A statistically significant difference between the observed and expected haplotype frequencies could

be due to one or more factors including significant inbreeding in the population group, strong selective pressure on the gene, sampling bias, and/or errors in the genotyping process. If large deviations from Hardy-Weinberg equilibrium are observed in an ethnogeographic group, the number of individuals in that group can be increased to see if the deviation is due to a sampling bias. If a larger sample size does not reduce the difference between observed and expected haplotype pair frequencies, then one may wish to consider haplotyping the individual using a direct haplotyping method such as, for example, CLASPER System™ technology (U.S. Patent No. 5,866,404), single molecule dilution, or allele-specific long-range PCR (Michalotos-Beloin et al., *Nucleic Acids Res.* 24:4841-4843, 1996).

In one embodiment of this method for predicting a CCR5 haplotype pair for an individual, the assigning step involves performing the following analysis. First, each of the possible haplotype pairs is compared to the haplotype pairs in the reference population. Generally, only one of the haplotype pairs in the reference population matches a possible haplotype pair and that pair is assigned to the individual. Occasionally, only one haplotype represented in the reference haplotype pairs is consistent with a possible haplotype pair for an individual, and in such cases the individual is assigned a haplotype pair containing this known haplotype and a new haplotype derived by subtracting the known haplotype from the possible haplotype pair. Alternatively, the haplotype pair in an individual may be predicted from the individual's genotype for that gene using reported methods (e.g., Clark et al. 1990 *Mol Bio Evol* 7:111-22) or through a commercial haplotyping service such as offered by Gennaissance Pharmaceuticals, Inc. (New Haven, CT). In rare cases, either no haplotypes in the reference population are consistent with the possible haplotype pairs, or alternatively, multiple reference haplotype pairs are consistent with the possible haplotype pairs. In such cases, the individual is preferably haplotyped using a direct molecular haplotyping method such as, for example, CLASPER System™ technology (U.S. Patent No. 5,866,404), SMD, or allele-specific long-range PCR (Michalotos-Beloin et al., *supra*).

The invention also provides a method for determining the frequency of a CCR5 genotype, haplotype, or haplotype pair in a population. The method comprises, for each member of the population, determining the genotype or the haplotype pair for the novel CCR5 polymorphic sites described herein, and calculating the frequency any particular genotype, haplotype, or haplotype pair is found in the population. The population may be a reference population, a family population, a same sex population, a population group, or a trait population (e.g., a group of individuals exhibiting a trait of interest such as a medical condition or response to a therapeutic treatment).

In another aspect of the invention, frequency data for CCR5 genotypes, haplotypes, and/or haplotype pairs are determined in a reference population and used in a method for identifying an association between a trait and a CCR5 genotype, haplotype, or haplotype pair. The trait may be any detectable phenotype, including but not limited to susceptibility to a disease or response to a treatment. The method involves obtaining data on the frequency of the genotype(s), haplotype(s), or haplotype pair(s) of interest in a reference population as well as in a population exhibiting the trait. Frequency data for one or both of the reference and trait populations may be obtained by genotyping or

haplotyping each individual in the populations using one of the methods described above. The haplotypes for the trait population may be determined directly or, alternatively, by the predictive genotype to haplotype approach described above. In another embodiment, the frequency data for the reference and/or trait populations is obtained by accessing previously determined frequency data, which may be in written or electronic form. For example, the frequency data may be present in a database that is accessible by a computer. Once the frequency data is obtained, the frequencies of the genotype(s), haplotype(s), or haplotype pair(s) of interest in the reference and trait populations are compared. In a preferred embodiment, the frequencies of all genotypes, haplotypes, and/or haplotype pairs observed in the populations are compared. If a particular CCR5 genotype, haplotype, or haplotype pair is more frequent in the trait population than in the reference population at a statistically significant amount, then the trait is predicted to be associated with that CCR5 genotype, haplotype, or haplotype pair. Preferably, the CCR5 genotype, haplotype, or haplotype pair being compared in the trait and reference populations is selected from the full-genotypes and full-haplotypes shown in Tables 4 and 5, or from sub-genotypes and sub-haplotypes derived from these genotypes and haplotypes.

In a preferred embodiment of the method, the trait of interest is a clinical response exhibited by a patient to some therapeutic treatment, for example, response to a drug targeting CCR5 or response to a therapeutic treatment for a medical condition. As used herein, "medical condition" includes but is not limited to any condition or disease manifested as one or more physical and/or psychological symptoms for which treatment is desirable, and includes previously and newly identified diseases and other disorders. As used herein the term "clinical response" means any or all of the following: a quantitative measure of the response, no response, and adverse response (i.e., side effects).

In order to deduce a correlation between clinical response to a treatment and a CCR5 genotype, haplotype, or haplotype pair, it is necessary to obtain data on the clinical responses exhibited by a population of individuals who received the treatment, hereinafter the "clinical population". This clinical data may be obtained by analyzing the results of a clinical trial that has already been run and/or the clinical data may be obtained by designing and carrying out one or more new clinical trials. As used herein, the term "clinical trial" means any research study designed to collect clinical data on responses to a particular treatment, and includes but is not limited to phase I, phase II and phase III clinical trials. Standard methods are used to define the patient population and to enroll subjects.

It is preferred that the individuals included in the clinical population have been graded for the existence of the medical condition of interest. This is important in cases where the symptom(s) being presented by the patients can be caused by more than one underlying condition, and where treatment of the underlying conditions are not the same. An example of this would be where patients experience breathing difficulties that are due to either asthma or respiratory infections. If both sets were treated with an asthma medication, there would be a spurious group of apparent non-responders that did not actually have asthma. These people would affect the ability to detect any correlation between haplotype and treatment outcome. This grading of potential patients could employ a standard physical exam or one or more lab tests. Alternatively, grading of patients could use haplotyping for situations

where there is a strong correlation between haplotype pair and disease susceptibility or severity.

The therapeutic treatment of interest is administered to each individual in the trial population and each individual's response to the treatment is measured using one or more predetermined criteria. It is contemplated that in many cases, the trial population will exhibit a range of responses and that the investigator will choose the number of responder groups (e.g., low, medium, high) made up by the various responses. In addition, the CCR5 gene for each individual in the trial population is genotyped and/or haplotyped, which may be done before or after administering the treatment.

After both the clinical and polymorphism data have been obtained, correlations between individual response and CCR5 genotype or haplotype content are created. Correlations may be produced in several ways. In one method, individuals are grouped by their CCR5 genotype or haplotype (or haplotype pair) (also referred to as a polymorphism group), and then the averages and standard deviations of clinical responses exhibited by the members of each polymorphism group are calculated.

These results are then analyzed to determine if any observed variation in clinical response between polymorphism groups is statistically significant. Statistical analysis methods which may be used are described in L.D. Fisher and G. vanBelle, "Biostatistics: A Methodology for the Health Sciences", Wiley-Interscience (New York) 1993. This analysis may also include a regression calculation of which polymorphic sites in the CCR5 gene give the most significant contribution to the differences in phenotype. One regression model useful in the invention is described in PCT Application Serial No. PCT/US00/17540, entitled "Methods for Obtaining and Using Haplotype Data".

A second method for finding correlations between CCR5 haplotype content and clinical responses uses predictive models based on error-minimizing optimization algorithms. One of many possible optimization algorithms is a genetic algorithm (R. Judson, "Genetic Algorithms and Their Uses in Chemistry" in Reviews in Computational Chemistry, Vol. 10, pp. 1-73, K. B. Lipkowitz and D. B. Boyd, eds. (VCH Publishers, New York, 1997). Simulated annealing (Press et al., "Numerical Recipes in C: The Art of Scientific Computing", Cambridge University Press (Cambridge) 1992, Ch. 10), neural networks (E. Rich and K. Knight, "Artificial Intelligence", 2nd Edition (McGraw-Hill, New York, 1991, Ch. 18), standard gradient descent methods (Press et al., *supra*, Ch. 10), or other global or local optimization approaches (see discussion in Judson, *supra*) could also be used. Preferably, the correlation is found using a genetic algorithm approach as described in PCT Application Serial No. PCT/US00/17540.

Correlations may also be analyzed using analysis of variation (ANOVA) techniques to determine how much of the variation in the clinical data is explained by different subsets of the polymorphic sites in the CCR5 gene. As described in PCT Application Serial No. PCT/US00/17540, ANOVA is used to test hypotheses about whether a response variable is caused by or correlated with one or more traits or variables that can be measured (Fisher and vanBelle, *supra*, Ch. 10).

From the analyses described above, a mathematical model may be readily constructed by the

skilled artisan that predicts clinical response as a function of CCR5 genotype or haplotype content. Preferably, the model is validated in one or more follow-up clinical trials designed to test the model.

The identification of an association between a clinical response and a genotype or haplotype (or haplotype pair) for the CCR5 gene may be the basis for designing a diagnostic method to determine those individuals who will or will not respond to the treatment, or alternatively, will respond at a lower level and thus may require more treatment, i.e., a greater dose of a drug. The diagnostic method may take one of several forms: for example, a direct DNA test (i.e., genotyping or haplotyping one or more of the polymorphic sites in the CCR5 gene), a serological test, or a physical exam measurement. The only requirement is that there be a good correlation between the diagnostic test results and the underlying CCR5 genotype or haplotype that is in turn correlated with the clinical response. In a preferred embodiment, this diagnostic method uses the predictive haplotyping method described above.

In another embodiment, the invention provides an isolated polynucleotide comprising a polymorphic variant of the CCR5 gene or a fragment of the gene which contains at least one of the novel polymorphic sites described herein. The nucleotide sequence of a variant CCR5 gene is identical to the reference genomic sequence for those portions of the gene examined, as described in the Examples below, except that it comprises a different nucleotide at one or more of the novel polymorphic sites PS6, PS7, PS10 and PS11, and may also comprise one or more additional polymorphisms selected from the group consisting of adenine at PS1, cytosine at PS2, thymine at PS3, adenine at PS4, thymine at PS5, adenine at PS8, cytosine at PS9 and adenine at PS12. Similarly, the nucleotide sequence of a variant fragment of the CCR5 gene is identical to the corresponding portion of the reference sequence except for having a different nucleotide at one or more of the novel polymorphic sites described herein. Thus, the invention specifically does not include polynucleotides comprising a nucleotide sequence identical to the reference sequence of the CCR5 gene, which is defined by haplotype 1, (or other reported CCR5 sequences) or to portions of the reference sequence (or other reported CCR5 sequences), except for genotyping oligonucleotides as described below.

The location of a polymorphism in a variant gene or fragment is identified by aligning its sequence against SEQ ID NO:1. The polymorphism is selected from the group consisting of cytosine at PS6, thymine at PS7, thymine at PS10 and guanine at PS11. In a preferred embodiment, the polymorphic variant comprises a naturally-occurring isogene of the CCR5 gene which is defined by any one of haplotypes 2-11 shown in Table 5 below.

Polymorphic variants of the invention may be prepared by isolating a clone containing the CCR5 gene from a human genomic library. The clone may be sequenced to determine the identity of the nucleotides at the novel polymorphic sites described herein. Any particular variant claimed herein could be prepared from this clone by performing *in vitro* mutagenesis using procedures well-known in the art.

CCR5 isogenes may be isolated using any method that allows separation of the two "copies" of the CCR5 gene present in an individual, which, as readily understood by the skilled artisan, may be

the same allele or different alleles. Separation methods include targeted *in vivo* cloning (TIVC) in yeast as described in WO 98/01573, U.S. Patent No. 5,866,404, and U.S. Patent No. 5,972,614. Another method, which is described in U.S. Patent No. 5,972,614, uses an allele specific oligonucleotide in combination with primer extension and exonuclease degradation to generate hemizygous DNA targets. Yet other methods are single molecule dilution (SMD) as described in Ruaño et al., *Proc. Natl. Acad. Sci.* 87:6296-6300, 1990; and allele specific PCR (Ruaño et al., 1989, *supra*; Ruaño et al., 1991, *supra*; Michalatos-Beloin et al., *supra*).

The invention also provides CCR5 genome anthologies, which are collections of CCR5 isogenes found in a given population. The population may be any group of at least two individuals, including but not limited to a reference population, a population group, a family population, a clinical population, and a same sex population. A CCR5 genome anthology may comprise individual CCR5 isogenes stored in separate containers such as microtest tubes, separate wells of a microtitre plate and the like. Alternatively, two or more groups of the CCR5 isogenes in the anthology may be stored in separate containers. Individual isogenes or groups of isogenes in a genome anthology may be stored in any convenient and stable form, including but not limited to in buffered solutions, as DNA precipitates, freeze-dried preparations and the like. A preferred CCR5 genome anthology of the invention comprises a set of isogenes defined by the haplotypes shown in Table 5 below.

An isolated polynucleotide containing a polymorphic variant nucleotide sequence of the invention may be operably linked to one or more expression regulatory elements in a recombinant expression vector capable of being propagated and expressing the encoded CCR5 protein in a prokaryotic or a eukaryotic host cell. Examples of expression regulatory elements which may be used include, but are not limited to, the lac system, operator and promoter regions of phage lambda, yeast promoters, and promoters derived from vaccinia virus, adenovirus, retroviruses, or SV40. Other regulatory elements include, but are not limited to, appropriate leader sequences, termination codons, polyadenylation signals, and other sequences required for the appropriate transcription and subsequent translation of the nucleic acid sequence in a given host cell. Of course, the correct combinations of expression regulatory elements will depend on the host system used. In addition, it is understood that the expression vector contains any additional elements necessary for its transfer to and subsequent replication in the host cell. Examples of such elements include, but are not limited to, origins of replication and selectable markers. Such expression vectors are commercially available or are readily constructed using methods known to those in the art (e.g., F. Ausubel et al., 1987, in "Current Protocols in Molecular Biology", John Wiley and Sons, New York, New York). Host cells which may be used to express the variant CCR5 sequences of the invention include, but are not limited to, eukaryotic and mammalian cells, such as animal, plant, insect and yeast cells, and prokaryotic cells, such as *E. coli*, or algal cells as known in the art. The recombinant expression vector may be introduced into the host cell using any method known to those in the art including, but not limited to, microinjection, electroporation, particle bombardment, transduction, and transfection using DEAE-dextran, lipofection, or calcium phosphate (see e.g., Sambrook et al. (1989) in "Molecular Cloning. A

Laboratory Manual", Cold Spring Harbor Press, Plainview, New York). In a preferred aspect, eukaryotic expression vectors that function in eukaryotic cells, and preferably mammalian cells, are used. Non-limiting examples of such vectors include vaccinia virus vectors, adenovirus vectors, herpes virus vectors, and baculovirus transfer vectors. Preferred eukaryotic cell lines include COS cells, CHO cells, HeLa cells, NIH/3T3 cells, and embryonic stem cells (Thomson, J. A. et al., 1998 *Science* 282:1145-1147). Particularly preferred host cells are mammalian cells.

As will be readily recognized by the skilled artisan, expression of polymorphic variants of the CCR5 gene will produce CCR5 mRNAs varying from each other at any polymorphic site retained in the spliced and processed mRNA molecules. These mRNAs can be used for the preparation of a CCR5 cDNA comprising a nucleotide sequence which is a polymorphic variant of the CCR5 reference coding sequence shown in Figure 2. Thus, the invention also provides CCR5 mRNAs and corresponding cDNAs which comprise a nucleotide sequence that is identical to SEQ ID NO: 2 (Fig. 2), or its corresponding RNA sequence, except for having one or more polymorphisms selected from the group consisting of thymine at a position corresponding to nucleotide 321 and guanine at a position corresponding to nucleotide 546, and may also comprise one or more additional polymorphisms selected from the group consisting of adenine at a position corresponding to nucleotide 164, cytosine at a position corresponding to nucleotide 225 and adenine at a position corresponding to nucleotide 668. A particularly preferred polymorphic cDNA variant comprises the coding sequence of a CCR5 isogene defined by haplotypes 2-11. Fragments of these variant mRNAs and cDNAs are included in the scope of the invention, provided they contain the novel polymorphisms described herein. The invention specifically excludes polynucleotides identical to previously identified and characterized CCR5 cDNAs and fragments thereof. Polynucleotides comprising a variant RNA or DNA sequence may be isolated from a biological sample using well-known molecular biological procedures or may be chemically synthesized.

As used herein, a polymorphic variant of a CCR5 gene fragment comprises at least one novel polymorphism identified herein and has a length of at least 10 nucleotides and may range up to the full length of the gene. Preferably, such fragments are between 100 and 3000 nucleotides in length, and more preferably between 200 and 2000 nucleotides in length, and most preferably between 500 and 1000 nucleotides in length.

In describing the CCR5 polymorphic sites identified herein, reference is made to the sense strand of the gene for convenience. However, as recognized by the skilled artisan, nucleic acid molecules containing the CCR5 gene may be complementary double stranded molecules and thus reference to a particular site on the sense strand refers as well to the corresponding site on the complementary antisense strand. Thus, reference may be made to the same polymorphic site on either strand and an oligonucleotide may be designed to hybridize specifically to either strand at a target region containing the polymorphic site. Thus, the invention also includes single-stranded polynucleotides which are complementary to the sense strand of the CCR5 genomic variants described herein.

Polynucleotides comprising a polymorphic gene variant or fragment may be useful for therapeutic purposes. For example, where a patient could benefit from expression, or increased expression, of a particular CCR5 protein isoform, an expression vector encoding the isoform may be administered to the patient. The patient may be one who lacks the CCR5 isogene encoding that isoform or may already have at least one copy of that isogene.

In other situations, it may be desirable to decrease or block expression of a particular CCR5 isogene. Expression of a CCR5 isogene may be turned off by transforming a targeted organ, tissue or cell population with an expression vector that expresses high levels of untranslatable mRNA for the isogene. Alternatively, oligonucleotides directed against the regulatory regions (e.g., promoter, introns, enhancers, 3' untranslated region) of the isogene may block transcription. Oligonucleotides targeting the transcription initiation site, e.g., between positions -10 and +10 from the start site are preferred. Similarly, inhibition of transcription can be achieved using oligonucleotides that base-pair with region(s) of the isogene DNA to form triplex DNA (see e.g., Gee et al. in Huber, B.E. and B.I. Carr, *Molecular and Immunologic Approaches*, Futura Publishing Co., Mt. Kisco, N.Y., 1994). Antisense oligonucleotides may also be designed to block translation of CCR5 mRNA transcribed from a particular isogene. It is also contemplated that ribozymes may be designed that can catalyze the specific cleavage of CCR5 mRNA transcribed from a particular isogene.

The oligonucleotides may be delivered to a target cell or tissue by expression from a vector introduced into the cell or tissue *in vivo* or *ex vivo*. Alternatively, the oligonucleotides may be formulated as a pharmaceutical composition for administration to the patient. Oligoribonucleotides and/or oligodeoxynucleotides intended for use as antisense oligonucleotides may be modified to increase stability and half-life. Possible modifications include, but are not limited to phosphorothioate or 2' O-methyl linkages, and the inclusion of nontraditional bases such as inosine and queosine, as well as acetyl-, methyl-, thio-, and similarly modified forms of adenine, cytosine, guanine, thymine, and uracil which are not as easily recognized by endogenous nucleases.

The invention also provides an isolated polypeptide comprising a polymorphic variant of the reference CCR5 amino acid sequence shown in Figure 3. The location of a variant amino acid in a CCR5 polypeptide or fragment of the invention is identified by aligning its sequence against SEQ ID NO:3 (Fig.3). A CCR5 protein variant of the invention comprises an amino acid sequence identical to SEQ ID NO:3 except for having leucine at a position corresponding to amino acid position 182, and may also comprise one or more additional variant amino acids selected from the group consisting of glutamine at a position corresponding to amino acid position 55 and glutamine at a position corresponding to amino acid position 223. The invention specifically excludes amino acid sequences identical to those previously identified for CCR5, including SEQ ID NO:3, and previously described fragments thereof. CCR5 protein variants included within the invention comprise all amino acid sequences based on SEQ ID NO:3 and having the combination of amino acid variations described in Table 2 below. In preferred embodiments, a CCR5 protein variant of the invention is encoded by an isogene defined by one of the observed haplotypes shown in Table 5.

Table 2. Novel Polymorphic Variants of CCR5

Polymorphic Variant Number	Amino Acid Position and Identities		
	55	182	223
1	L	F	Q
2	L	L	R
3	L	L	Q
4	Q	F	Q
5	Q	L	R
6	Q	L	Q

The invention also includes CCR5 peptide variants, which are any fragments of a CCR5 protein variant that contain leucine at a position corresponding to amino acid position 182. A CCR5 peptide variant is at least 6 amino acids in length and is preferably any number between 6 and 30 amino acids long, more preferably between 10 and 25, and most preferably between 15 and 20 amino acids long. Such CCR5 peptide variants may be useful as antigens to generate antibodies specific for one of the above CCR5 isoforms. In addition, the CCR5 peptide variants may be useful in drug screening assays.

A CCR5 variant protein or peptide of the invention may be prepared by chemical synthesis or by expressing one of the variant CCR5 genomic and cDNA sequences as described above. Alternatively, the CCR5 protein variant may be isolated from a biological sample of an individual having a CCR5 isogene which encodes the variant protein. Where the sample contains two different CCR5 isoforms (i.e., the individual has different CCR5 isogenes), a particular CCR5 isoform of the invention can be isolated by immunoaffinity chromatography using an antibody which specifically binds to that particular CCR5 isoform but does not bind to the other CCR5 isoform.

The expressed or isolated CCR5 protein may be detected by methods known in the art, including Coomassie blue staining, silver staining, and Western blot analysis using antibodies specific for the isoform of the CCR5 protein as discussed further below. CCR5 variant proteins can be purified by standard protein purification procedures known in the art, including differential precipitation, molecular sieve chromatography, ion-exchange chromatography, isoelectric focusing, gel electrophoresis, affinity and immunoaffinity chromatography and the like. (Ausubel et. al., 1987, In Current Protocols in Molecular Biology John Wiley and Sons, New York, New York). In the case of immunoaffinity chromatography, antibodies specific for a particular polymorphic variant may be used.

A polymorphic variant CCR5 gene of the invention may also be fused in frame with a heterologous sequence to encode a chimeric CCR5 protein. The non-CCR5 portion of the chimeric protein may be recognized by a commercially available antibody. In addition, the chimeric protein may also be engineered to contain a cleavage site located between the CCR5 and non-CCR5 portions so that the CCR5 protein may be cleaved and purified away from the non-CCR5 portion.

An additional embodiment of the invention relates to using a novel CCR5 protein isoform in

any of a variety of drug screening assays. Such screening assays may be performed to identify agents that bind specifically to all known CCR5 protein isoforms or to only a subset of one or more of these isoforms. The agents may be from chemical compound libraries, peptide libraries and the like. The CCR5 protein or peptide variant may be free in solution or affixed to a solid support. In one embodiment, high throughput screening of compounds for binding to a CCR5 variant may be accomplished using the method described in PCT application WO84/03565, in which large numbers of test compounds are synthesized on a solid substrate, such as plastic pins or some other surface, contacted with the CCR5 protein(s) of interest and then washed. Bound CCR5 protein(s) are then detected using methods well-known in the art.

In another embodiment, a novel CCR5 protein isoform may be used in assays to measure the binding affinities of one or more candidate drugs targeting the CCR5 protein.

In yet another embodiment, when a particular CCR5 haplotype or group of CCR5 haplotypes encodes a CCR5 protein variant with an amino acid sequence distinct from that of CCR5 protein isoforms encoded by other CCR5 haplotypes, then detection of that particular CCR5 haplotype or group of CCR5 haplotypes may be accomplished by detecting expression of the encoded CCR5 protein variant using any of the methods described herein or otherwise commonly known to the skilled artisan.

In another embodiment, the invention provides antibodies specific for and immunoreactive with one or more of the novel CCR5 variant proteins described herein. The antibodies may be either monoclonal or polyclonal in origin. The CCR5 protein or peptide variant used to generate the antibodies may be from natural or recombinant sources or produced by chemical synthesis using synthesis techniques known in the art. If the CCR5 protein variant is of insufficient size to be antigenic, it may be conjugated, complexed, or otherwise covalently linked to a carrier molecule to enhance the antigenicity of the peptide. Examples of carrier molecules, include, but are not limited to, albumins (e.g., human, bovine, fish, ovine), and keyhole limpet hemocyanin (Basic and Clinical Immunology, 1991, Eds. D.P. Stites, and A.I. Terr, Appleton and Lange, Norwalk Connecticut, San Mateo, California).

In one embodiment, an antibody specifically immunoreactive with one of the novel protein isoforms described herein is administered to an individual to neutralize activity of the CCR5 isoform expressed by that individual. The antibody may be formulated as a pharmaceutical composition which includes a pharmaceutically acceptable carrier.

Antibodies specific for and immunoreactive with one of the novel protein isoforms described herein may be used to immunoprecipitate the CCR5 protein variant from solution as well as react with CCR5 protein isoforms on Western or immunoblots of polyacrylamide gels on membrane supports or substrates. In another preferred embodiment, the antibodies will detect CCR5 protein isoforms in paraffin or frozen tissue sections, or in cells which have been fixed or unfixed and prepared on slides, coverslips, or the like, for use in immunocytochemical, immunohistochemical, and immunofluorescence techniques.

In another embodiment, an antibody specifically immunoreactive with one of the novel CCR5

protein variants described herein is used in immunoassays to detect this variant in biological samples. In this method, an antibody of the present invention is contacted with a biological sample and the formation of a complex between the CCR5 protein variant and the antibody is detected. As described, suitable immunoassays include radioimmunoassay, Western blot assay, immunofluorescent assay, enzyme linked immunoassay (ELISA), chemiluminescent assay, immunohistochemical assay, immunocytochemical assay, and the like (see, e.g., *Principles and Practice of Immunoassay*, 1991, Eds. Christopher P. Price and David J. Neoman, Stockton Press, New York, New York; *Current Protocols in Molecular Biology*, 1987, Eds. Ausubel et al., John Wiley and Sons, New York, New York). Standard techniques known in the art for ELISA are described in *Methods in Immunodiagnosis*, 2nd Ed., Eds. Rose and Bigazzi, John Wiley and Sons, New York 1980; and Campbell et al., 1984, *Methods in Immunology*, W.A. Benjamin, Inc.). Such assays may be direct, indirect, competitive, or noncompetitive as described in the art (see, e.g., *Principles and Practice of Immunoassay*, 1991, Eds. Christopher P. Price and David J. Neoman, Stockton Press, NY, NY; and Oellirich, M., 1984, *J. Clin. Chem. Clin. Biochem.*, 22:895-904). Proteins may be isolated from test specimens and biological samples by conventional methods, as described in *Current Protocols in Molecular Biology*, supra.

Exemplary antibody molecules for use in the detection and therapy methods of the present invention are intact immunoglobulin molecules, substantially intact immunoglobulin molecules, or those portions of immunoglobulin molecules that contain the antigen binding site. Polyclonal or monoclonal antibodies may be produced by methods conventionally known in the art (e.g., Kohler and Milstein, 1975, *Nature*, 256:495-497; Campbell *Monoclonal Antibody Technology, the Production and Characterization of Rodent and Human Hybridomas*, 1985, In: *Laboratory Techniques in Biochemistry and Molecular Biology*, Eds. Burdon et al., Volume 13, Elsevier Science Publishers, Amsterdam). The antibodies or antigen binding fragments thereof may also be produced by genetic engineering. The technology for expression of both heavy and light chain genes in *E. coli* is the subject of PCT patent applications, publication number WO 901443, WO 901443 and WO 9014424 and in Huse et al., 1989, *Science*, 246:1275-1281. The antibodies may also be humanized (e.g., Queen, C. et al. 1989 *Proc. Natl. Acad. Sci. USA* 86:10029).

Effect(s) of the polymorphisms identified herein on expression of CCR5 may be investigated by preparing recombinant cells and/or nonhuman recombinant organisms, preferably recombinant animals, containing a polymorphic variant of the CCR5 gene. As used herein, "expression" includes but is not limited to one or more of the following: transcription of the gene into precursor mRNA; splicing and other processing of the precursor mRNA to produce mature mRNA; mRNA stability; translation of the mature mRNA into CCR5 protein (including codon usage and tRNA availability); and glycosylation and/or other modifications of the translation product, if required for proper expression and function.

To prepare a recombinant cell of the invention, the desired CCR5 isogene may be introduced into the cell in a vector such that the isogene remains extrachromosomal. In such a situation, the gene will be expressed by the cell from the extrachromosomal location. In a preferred embodiment, the

CCR5 isogene is introduced into a cell in such a way that it recombines with the endogenous CCR5 gene present in the cell. Such recombination requires the occurrence of a double recombination event, thereby resulting in the desired CCR5 gene polymorphism. Vectors for the introduction of genes both for recombination and for extrachromosomal maintenance are known in the art, and any suitable vector or vector construct may be used in the invention. Methods such as electroporation, particle bombardment, calcium phosphate co-precipitation and viral transduction for introducing DNA into cells are known in the art; therefore, the choice of method may lie with the competence and preference of the skilled practitioner. Examples of cells into which the CCR5 isogene may be introduced include, but are not limited to, continuous culture cells, such as COS, NIH/3T3, and primary or culture cells of the relevant tissue type, i.e., they express the CCR5 isogene. Such recombinant cells can be used to compare the biological activities of the different protein variants.

Recombinant nonhuman organisms, i.e., transgenic animals, expressing a variant CCR5 gene are prepared using standard procedures known in the art. Preferably, a construct comprising the variant gene is introduced into a nonhuman animal or an ancestor of the animal at an embryonic stage, i.e., the one-cell stage, or generally not later than about the eight-cell stage. Transgenic animals carrying the constructs of the invention can be made by several methods known to those having skill in the art. One method involves transfecting into the embryo a retrovirus constructed to contain one or more insulator elements, a gene or genes of interest, and other components known to those skilled in the art to provide a complete shuttle vector harboring the insulated gene(s) as a transgene, see e.g., U.S. Patent No. 5,610,053. Another method involves directly injecting a transgene into the embryo. A third method involves the use of embryonic stem cells. Examples of animals into which the CCR5 isogenes may be introduced include, but are not limited to, mice, rats, other rodents, and nonhuman primates (see "The Introduction of Foreign Genes into Mice" and the cited references therein, In: Recombinant DNA, Eds. J.D. Watson, M. Gilman, J. Witkowski, and M. Zoller; W.H. Freeman and Company, New York, pages 254-272). Transgenic animals stably expressing a human CCR5 isogene and producing human CCR5 protein can be used as biological models for studying diseases related to abnormal CCR5 expression and/or activity, and for screening and assaying various candidate drugs, compounds, and treatment regimens to reduce the symptoms or effects of these diseases.

An additional embodiment of the invention relates to pharmaceutical compositions for treating disorders affected by expression or function of a novel CCR5 isogene described herein. The pharmaceutical composition may comprise any of the following active ingredients: a polynucleotide comprising one of these novel CCR5 isogenes; an antisense oligonucleotide directed against one of the novel CCR5 isogenes, a polynucleotide encoding such an antisense oligonucleotide, or another compound which inhibits expression of a novel CCR5 isogene described herein. Preferably, the composition contains the active ingredient in a therapeutically effective amount. By therapeutically effective amount is meant that one or more of the symptoms relating to disorders affected by expression or function of a novel CCR5 isogene is reduced and/or eliminated. The composition also comprises a pharmaceutically acceptable carrier, examples of which include, but are not limited to,

saline, buffered saline, dextrose, and water. Those skilled in the art may employ a formulation most suitable for the active ingredient, whether it is a polynucleotide, oligonucleotide, protein, peptide or small molecule antagonist. The pharmaceutical composition may be administered alone or in combination with at least one other agent, such as a stabilizing compound. Administration of the pharmaceutical composition may be by any number of routes including, but not limited to oral, intravenous, intramuscular, intra-arterial, intramedullary, intrathecal, intraventricular, intradermal, transdermal, subcutaneous, intraperitoneal, intranasal, enteral, topical, sublingual, or rectal. Further details on techniques for formulation and administration may be found in the latest edition of Remington's Pharmaceutical Sciences (Maack Publishing Co., Easton, PA).

For any composition, determination of the therapeutically effective dose of active ingredient and/or the appropriate route of administration is well within the capability of those skilled in the art. For example, the dose can be estimated initially either in cell culture assays or in animal models. The animal model may also be used to determine the appropriate concentration range and route of administration. Such information can then be used to determine useful doses and routes for administration in humans. The exact dosage will be determined by the practitioner, in light of factors relating to the patient requiring treatment, including but not limited to severity of the disease state, general health, age, weight and gender of the patient, diet, time and frequency of administration, other drugs being taken by the patient, and tolerance/response to the treatment.

Any or all analytical and mathematical operations involved in practicing the methods of the present invention may be implemented by a computer. In addition, the computer may execute a program that generates views (or screens) displayed on a display device and with which the user can interact to view and analyze large amounts of information relating to the CCR5 gene and its genomic variation, including chromosome location, gene structure, and gene family, gene expression data, polymorphism data, genetic sequence data, and clinical data population data (e.g., data on ethnogeographic origin, clinical responses, genotypes, and haplotypes for one or more populations). The CCR5 polymorphism data described herein may be stored as part of a relational database (e.g., an instance of an Oracle database or a set of ASCII flat files). These polymorphism data may be stored on the computer's hard drive or may, for example, be stored on a CD-ROM or on one or more other storage devices accessible by the computer. For example, the data may be stored on one or more databases in communication with the computer via a network.

Preferred embodiments of the invention are described in the following examples. Other embodiments within the scope of the claims herein will be apparent to one skilled in the art from consideration of the specification or practice of the invention as disclosed herein. It is intended that the specification, together with the examples, be considered exemplary only, with the scope and spirit of the invention being indicated by the claims which follow the examples.

EXAMPLES

The Examples herein are meant to exemplify the various aspects of carrying out the invention

and are not intended to limit the scope of the invention in any way. The Examples do not include detailed descriptions for conventional methods employed, such as in the performance of genomic DNA isolation, PCR and sequencing procedures. Such methods are well-known to those skilled in the art and are described in numerous publications, for example, Sambrook, Fritsch, and Maniatis, "Molecular Cloning: A Laboratory Manual", 2nd Edition, Cold Spring Harbor Laboratory Press, USA, (1989).

EXAMPLE 1

This example illustrates examination of various regions of the CCR5 gene for polymorphic sites.

Amplification of Target Regions

The following target regions of the CCR5 gene were amplified using PCR primer pairs. The primers used for each region are represented below by providing the nucleotide positions of their initial and final nucleotides, which correspond to positions in the indicated GenBank Accession Number.

PCR Primer Pairs

GenBank Acc No.	Fragment No.	Forward Primer	Reverse Primer (complement of)	PCR Product
U95626.1	Fragment 1	58995-59017	59596-59572	602 nt
	Fragment 2	59282-59305	59919-59896	638 nt
	Fragment 3	61268-61290	61887-61867	620 nt
	Fragment 4	61580-61600	62134-62112	555 nt
	Fragment 5	61833-61855	62451-62432	619 nt
	Fragment 6	62180-62199	62789-62767	610 nt
	Fragment 7	62333-62352	62999-62977	667 nt

These primer pairs were used in PCR reactions containing genomic DNA isolated from immortalized cell lines for each member of the Index Repository. The PCR reactions were carried out under the following conditions:

Reaction volume	= 10 µl
10 x Advantage 2 Polymerase reaction buffer (Clontech)	= 1 µl
100 ng of human genomic DNA	= 1 µl
10 mM dNTP	= 0.4 µl
Advantage 2 Polymerase enzyme mix (Clontech)	= 0.2 µl
Forward Primer (10 µM)	= 0.4 µl
Reverse Primer (10 µM)	= 0.4 µl
Water	= 6.6 µl

Amplification profile:

97°C - 2 min. 1 cycle

97°C - 15 sec.	}	10 cycles
70°C - 45 sec.		
72°C - 45 sec.		

97°C - 15 sec.
 64°C - 45 sec. } 35 cycles
 72°C - 45 sec.

Sequencing of PCR Products

The PCR products were purified using a Whatman/Polyfiltronics 100 µl 384 well unfilter plate essentially according to the manufacturers protocol. The purified DNA was eluted in 50 µl of distilled water. Sequencing reactions were set up using Applied Biosystems Big Dye Terminator chemistry essentially according to the manufacturers protocol. The purified PCR products were sequenced in both directions using the primer sets described previously or those represented below by the nucleotide positions of their initial and final nucleotides, which correspond to positions in the indicated GenBank Accession Number. Reaction products were purified by isopropanol precipitation, and run on an Applied Biosystems 3700 DNA Analyzer.

Sequencing Primer Pairs

GenBank Acc No.	Fragment No.	Forward Primer	Reverse Primer (complement of)
U95626.1	Fragment 1	59025-59044	59556-59537
	Fragment 2	59328-59347	59868-59847
	Fragment 3	61323-61344	61855-61834
	Fragment 4	61615-61634	62111-62092
	Fragment 5	61908-61927	62386-62366
	Fragment 6	62244-62263	62740-62721
	Fragment 7	62462-62481	62963-62946

Analysis of Sequences for Polymorphic Sites

Sequences were analyzed for the presence of polymorphisms using the Polyphred program (Nickerson et al., *Nucleic Acids Res.* 14:2745-2751, 1997). The presence of a polymorphism was confirmed on both strands. The polymorphisms and their locations in the CCR5 gene are listed in Table 3 below.

Table 3. Polymorphic Sites Identified in the CCR5 Gene

Polymorphic Site Number	PolyId ^a	Nucleotide Position	Reference Allele	Variant Allele
PS1 ^R	13383	59029(Acc#U95626.1)	G	A
PS2 ^R	13388	59353(Acc#U95626.1)	T	C
PS3 ^R	13389	59356(Acc#U95626.1)	C	T
PS4 ^R	13390	59402(Acc#U95626.1)	G	A
PS5 ^R	13453	59653(Acc#U95626.1)	C	T
PS6	13456	59802(Acc#U95626.1)	A	C
PS7	13405	61370(Acc#U95626.1)	G	T
PS8 ^R	13416	61646(Acc#U95626.1)	T	A
PS9 ^R	13418	61707(Acc#U95626.1)	T	C
PS10	13420	61803(Acc#U95626.1)	C	T
PS11	13424	62028(Acc#U95626.1)	T	G
PS12 ^R	13427	62150(Acc#U95626.1)	G	A

^Rpreviously identified in the literature

^aPolyId is a unique identifier assigned to each PS by Genaissance Pharmaceuticals, Inc.

EXAMPLE 2

This example illustrates analysis of the CCR5 polymorphisms identified in the Index Repository for human genotypes and haplotypes.

The different genotypes containing these polymorphisms that were observed in the reference population are shown in Table 4 below, with the haplotype pair indicating the combination of haplotypes determined for the individual using the haplotype derivation protocol described below. In Table 4, homozygous positions are indicated by one nucleotide and heterozygous positions are indicated by two nucleotides. Missing nucleotides in any given genotype in Table 4 were inferred based on linkage disequilibrium and/or Mendelian inheritance.

Table 4 Genotypes and Haplotype Pairs Observed for the CCR5 Gene													
Genotype Number	Polymorphic Sites												HAP PAIR
	PS1	PS2	PS3	PS4	PS5	PS6	PS7	PS8	PS9	PS10	PS11	PS12	
1	G	T	C	G	C	A	G	T	T	C	T	G	1 1
2	GA	TC	C	AG	C	A	G	T	T	C	T	G	1 2
3	GA	TC	C	AG	OT	A	G	T	T	C	T	G	1 3
4	G	T	C	AG	C	AC	GT	T	T	C	T	G	1 4
5	G	T	OT	AG	C	A	G	T	T	C	T	G	1 5
6	G	T	C	AG	C	A	G	T	T	C	T	G	1 6
7	GA	TC	C	AG	C	A	G	TA	T	C	T	G	1 7
8	-	TC	C	AG	OT	A	-	T	T	C	T	GA	1 11
9	A	C	C	A	C	A	G	T	T	C	T	G	2 2
10	A	C	C	A	OT	A	G	T	T	C	T	G	2 3
11	GA	TC	C	A	C	AC	GT	T	T	C	T	G	2 4
12	GA	TC	C	A	C	A	G	T	T	C	T	G	2 6
13	A	C	C	A	C	A	G	TA	T	C	T	G	2 7
14	A	C	C	A	OT	A	G	T	T	OT	T	G	2 9
15	A	C	C	A	OT	A	G	T	T	C	TG	G	2 10
16	A	C	C	A	T	A	G	T	T	C	T	G	3 3
17	GA	TC	C	A	OT	AC	GT	T	T	C	T	G	3 4
18	GA	TC	C	A	OT	A	G	T	T	C	T	G	3 6
19	G	T	OT	A	C	AC	GT	T	T	C	T	G	4 5
20	GA	TC	C	A	C	AC	GT	TA	T	C	T	G	4 7
21	G	T	OT	A	C	AC	GT	T	TC	C	T	G	4 8
22	G	T	T	A	C	A	G	T	T	C	T	G	5 5
23	G	T	T	A	C	A	G	T	TC	C	T	G	5 8
24	G	T	C	A	C	A	G	T	T	C	T	G	6 6

The haplotype pairs shown in Table 4 were estimated from the unphased genotypes using a computer-implemented extension of Clark's algorithm (Clark, A.G. 1990 *Mol Bio Evol* 7, 111-122) for

Table 5, Haplotypes Observed in the CCR5 Gene												
Haplotype Number	Polymorphic Sites											
	PS1	PS2	PS3	PS4	PS5	PS6	PS7	PS8	PS9	PS10	PS11	PS12
1.	G	T	C	G	C	A	G	T	T	C	T	G
2	A	C	C	A	C	A	G	T	T	C	T	G
3	A	C	C	A	T	A	G	T	T	C	T	G
4	G	T	C	A	C	C	T	T	T	C	T	G
5	G	T	T	A	C	A	G	T	T	C	T	G
6	G	T	C	A	C	A	G	T	T	C	T	G
7	A	C	C	A	C	A	G	A	T	C	T	G
8	G	T	T	A	C	A	G	T	C	C	T	G
9	A	C	C	A	T	A	G	T	T	T	T	G
10	A	C	C	A	T	A	G	T	T	C	G	G
11	A	C	C	A	T	A	G	T	T	C	T	A

The number of chromosomes for which a given CCR5 haplotype occurs in each population within the index repository is shown in Table 6.

Table 6. Haplotypes Identified in the CCR5 Gene

Haplotype	AF	AS	CA	HL	OT	Total
1	11	24	12	12	0	59
2	4	9	20	11	3	47
3	3	6	3	7	2	21
4	10	0	0	1	0	11
5	10	0	0	0	0	10
6	0	0	4	3	0	9
7	0	0	2	2	0	4
8	2	0	0	0	0	2
9	0	0	0	0	1	1
10	0	0	1	0	0	1
11	0	1	0	0	0	1

The number of individuals for which a given CCR5 haplotype occurs in each population is shown in Table 7 below.

Table 7. Frequencies of Observed Haplotype Pairs in the CCR5 Gene

Haplotype Pair	AF	AS	CA	HL	AM	Total
1,1	1	7	2	4	0	14
1,2	3	6	7	2	0	18
1,3	1	3	1	0	0	5
1,4	7	0	0	0	0	7
1,5	3	0	0	0	0	3
1,6	0	0	2	2	0	4
1,7	0	0	1	0	0	1
1,11	0	1	0	0	0	1
2,2	0	1	4	1	0	6
2,3	1	1	2	6	2	12
2,4	2	0	0	0	0	2
2,6	0	0	5	0	0	5
2,7	0	0	1	1	0	2
2,9	0	0	0	0	1	1
2,10	0	0	1	0	0	1
3,3	0	1	0	0	0	1
3,4	1	0	0	0	0	1
3,6	0	0	0	1	0	1
4,5	2	0	0	0	0	2
4,7	0	0	0	1	0	1
4,8	1	0	0	0	0	1
5,5	2	0	0	0	0	2
5,8	1	0	0	0	0	1
6,6	0	0	1	0	0	1

In view of the above, it will be seen that the several advantages of the invention are achieved and other advantageous results attained.

As various changes could be made in the above methods and compositions without departing from the scope of the invention, it is intended that all matter contained in the above description and shown in the accompanying drawings shall be interpreted as illustrative and not in a limiting sense.

All references cited in this specification, including patents and patent applications, are hereby incorporated in their entirety by reference. The discussion of references herein is intended merely to summarize the assertions made by their authors and no admission is made that any reference constitutes prior art. Applicants reserve the right to challenge the accuracy and pertinency of the cited references.

What is Claimed is:

1. A method for haplotyping the chemokine (C-C motif) receptor 5 (CCR5) gene of an individual which comprises determining whether the individual has one of the CCR5 haplotypes shown in Table 5 or one of the haplotype pairs shown in Table 4.
2. The method of claim 1, wherein the determining step comprises identifying the phased sequence of nucleotides present at each of PS1-PS12 on at least one copy of the individual's CCR5 gene.
3. The method of claim 1, wherein the determining step comprises identifying the phased sequence of nucleotides present at each of PS1-PS12 on both copies of the individual's CCR5 gene.
4. A method for genotyping the chemokine (C-C motif) receptor 5 (CCR5) gene of an individual, comprising determining for the two copies of the CCR5 gene present in the individual the identity of the nucleotide pair at one or more polymorphic sites selected from the group consisting of PS6, PS7, PS10 and PS11.
5. The method of claim 4, wherein the determining step comprises:
 - (a) isolating from the individual a nucleic acid mixture comprising both copies of the CCR5 gene, or a fragment thereof, that are present in the individual;
 - (b) amplifying from the nucleic acid mixture a target region containing the selected polymorphic site;
 - (c) hybridizing a primer extension oligonucleotide to one allele of the amplified target region;
 - (d) performing a nucleic acid template-dependent, primer extension reaction on the hybridized genotyping oligonucleotide in the presence of at least two different terminators of the reaction, wherein said terminators are complementary to the alternative nucleotides present at the selected polymorphic site; and
 - (e) detecting the presence and identity of the terminator in the extended genotyping oligonucleotide.
6. The method of claim 4, which comprises determining for the two copies of the CCR5 gene present in the individual the identity of the nucleotide pair at each of PS1-PS12.
7. A method for haplotyping the chemokine (C-C motif) receptor 5 (CCR5) gene of an individual which comprises determining, for one copy of the CCR5 gene present in the individual, the identity of the nucleotide at two or more polymorphic sites selected from the group consisting of PS6, PS7, PS10 and PS11.
8. The method of claim 7, further comprising determining the identity of the nucleotide at one or more polymorphic sites selected from the group consisting of PS1, PS2, PS3, PS4, PS5, PS8, PS9 and PS12.
9. The method of claim 7, wherein the determining step comprises:
 - (a) isolating from the individual a nucleic acid sample containing only one of the two copies of the CCR5 gene, or a fragment thereof, that is present in the individual;

- (b) amplifying from the nucleic acid molecule a target region containing the selected polymorphic site;
- (c) hybridizing a primer extension oligonucleotide to one allele of the amplified target region;
- (d) performing a nucleic acid template-dependent, primer extension reaction on the hybridized genotyping oligonucleotide in the presence of at least two different terminators of the reaction, wherein said terminators are complementary to the alternative nucleotides present at the selected polymorphic site; and
- (e) detecting the presence and identity of the terminator in the extended genotyping oligonucleotide.
10. A method for predicting a haplotype pair for the chemokine (C-C motif) receptor 5 (CCR5) gene of an individual comprising:
- (a) identifying a CCR5 genotype for the individual, wherein the genotype comprises the nucleotide pair at two or more polymorphic sites selected from the group consisting of PS6, PS7, PS10 and PS11;
- (b) enumerating all possible haplotype pairs which are consistent with the genotype;
- (c) comparing the possible haplotype pairs to the data in Table 4; and
- (d) assigning a haplotype pair to the individual that is consistent with the data.
11. The method of claim 10, wherein the identified genotype of the individual comprises the nucleotide pair at each of PS1-PS12.
12. A method for identifying an association between a trait and at least one haplotype or haplotype pair of the chemokine (C-C motif) receptor 5 (CCR5) gene which comprises comparing the frequency of the haplotype or haplotype pair in a population exhibiting the trait with the frequency of the haplotype or haplotype pair in a reference population, wherein the haplotype is selected from haplotypes 1-11 shown in Table 5 and the haplotype pair is selected from the haplotype pairs shown in Table 4, wherein a higher frequency of the haplotype or haplotype pair in the trait population than in the reference population indicates the trait is associated with the haplotype or haplotype pair.
13. The method of claim 12, wherein the trait is a clinical response to a drug targeting CCR5.
14. A composition comprising at least one genotyping oligonucleotide for detecting a polymorphism in the chemokine (C-C motif) receptor 5 (CCR5) gene at a polymorphic site selected from the group consisting of PS6, PS7, PS10 and PS11.
15. The composition of claim 14, wherein the genotyping oligonucleotide is an allele-specific oligonucleotide that specifically hybridizes to an allele of the CCR5 gene at a region containing the polymorphic site.
16. The composition of claim 15, wherein the allele-specific oligonucleotide comprises a nucleotide sequence selected from the group consisting of SEQ ID NOS:4-7, the complements of SEQ ID NOS:4-7, and SEQ ID NOS:8-15.

17. The composition of claim 14, wherein the genotyping oligonucleotide is a primer-extension oligonucleotide.
18. The composition of claim 17, wherein the primer extension oligonucleotide comprises a nucleotide sequence selected from the group consisting of SEQ ID NOS:16-23.
19. A kit for genotyping the CCR5 gene of an individual, which comprises a set of oligonucleotides designed to genotype each of PS6, PS7, PS10 and PS11.
20. The kit of claim 19, which further comprises oligonucleotides designed to genotype each of PS1, PS2, PS3, PS4, PS5, PS8, PS9 and PS12.
21. An isolated polynucleotide comprising a nucleotide sequence selected from the group consisting of:
 - (a) a first nucleotide sequence which is a polymorphic variant of a reference sequence for the chemokine (C-C motif) receptor 5 (CCR5) gene or a fragment thereof, wherein the reference sequence comprises SEQ ID NO:1 and the polymorphic variant comprises a CCR5 isogene defined by a haplotype selected from the group consisting of haplotypes 1-11 in Table 5; and
 - (b) a second nucleotide sequence which is complementary to the first nucleotide sequence.
22. The isolated polynucleotide of claim 21, which is a DNA molecule and comprises both the first and second nucleotide sequences and further comprises expression regulatory elements operably linked to the first nucleotide sequence.
23. A recombinant nonhuman organism transformed or transfected with the isolated polynucleotide of claim 21, wherein the organism expresses a CCR5 protein encoded by the first nucleotide sequence.
24. The recombinant organism of claim 23, which is a nonhuman transgenic animal.
25. The isolated polynucleotide of claim 21, wherein the first nucleotide sequence is a polymorphic variant of a fragment of the CCR5 gene, the fragment comprising one or more polymorphisms selected from the group consisting of cytosine at PS6, thymine at PS7, thymine at PS10 and guanine at PS11.
26. An isolated polynucleotide comprising a nucleotide sequence which is a polymorphic variant of a reference sequence for the CCR5 cDNA or a fragment thereof, wherein the reference sequence comprises SEQ ID NO:2 and the polymorphic variant comprises the coding sequence of a CCR5 isogene defined by one of the haplotypes shown in Table 5.
27. A recombinant nonhuman organism transformed or transfected with the isolated polynucleotide of claim 26, wherein the organism expresses a chemokine (C-C motif) receptor 5 (CCR5) protein encoded by the polymorphic variant sequence.
28. The recombinant organism of claim 27, which is a nonhuman transgenic animal.
29. An isolated polypeptide comprising an amino acid sequence which is a polymorphic variant of a reference sequence for the CCR5 protein or a fragment thereof, wherein the reference sequence

comprises SEQ ID NO:3 and the polymorphic variant is encoded by an isogene defined by one of the haplotypes shown in Table 5.

30. An isolated antibody specific for and immunoreactive with the isolated polypeptide of claim 29.
31. A method for screening for drugs targeting the isolated polypeptide of claim 29 which comprises contacting the CCR5 polymorphic variant with a candidate agent and assaying for binding activity.
32. A computer system for storing and analyzing polymorphism data for the chemokine (C-C motif) receptor 5 gene, comprising:
 - (a) a central processing unit (CPU);
 - (b) a communication interface;
 - (c) a display device;
 - 5 (d) an input device; and
 - (e) a database containing the polymorphism data;wherein the polymorphism data comprises the genotypes and haplotype pairs shown in Table 4 and the haplotypes shown in Table 5.
33. A genome anthology for the chemokine (C-C motif) receptor 5 (CCR5) gene which comprises CCR5 isogenes defined by any one of haplotypes 1-11 shown in Table 5.

POLYMORPHISMS IN THE CCR5 GENE (Accession No. U95626.1)

GTATGTCTTC	ACATCCTCCA	ATACCCTAAT	GACAGACAAA	CAGAACATGG	
CAAAGCCTCA	GCTCTGCATG	GTGAAAGTAA	GAACCAGCAA	TTGCCACAAA	57500
CAGAAATACA	GTGTTGGTCC	GGCAGCCTCC	GGGGGTCTG	CACAAGTGGA	
TTACCAGTGA	ATACAAGGCT	ATCTATCTTT	CGAAAAACCA	ACGTTGTATT	57600
TATGCTATTT	TCTATAAAAT	TTTATATTAA	TTTATTTGTT	ACCTATTTTT	
GAACTCTTTC	AAAAGCACAC	TTTATATFTC	CCTGCTTAAA	CAGTCCCCCG	57700
AGGGTGGGTG	CCCAAAAGGC	TCTACACTTG	TTATCATTC	CTCTCCACCA	
CAGGCATATT	GAGTAAGTTT	GTATTTGGGT	TTTTTTAAAA	CCTCCACTCT	57800
ACAGTTAAGA	AAACTAAGGC	ACAGAGCTTC	AATAATTTGG	TCAGAGCCAA	
GTAGCAGTAA	TGAAGCTGGA	GGTTAAACCC	AGCAGCATGA	CTGCAGTTCT	57900
TAATCAATGC	CTTTTGAATT	GCACATATGG	GATGAAGTAG	AACATTTTCT	
CGATGATTCT	CTGTCCTTGT	TATGATTATG	TTACTGAGCT	CTGTTGTAGC	58000
ACAGACATAT	GTCCCTATAT	GGGGCGGGGT	TGGGGGTGTC	TTGATCGCTG	
GGCTATTTCT	ATACTGTTCT	GGCTTTTCCC	AAGCAGTCAT	TTCTTTCTAT	58100
TCTCCAAGCA	CCAGCAATTA	GCTTTACCTT	TTCAGCTTCT	AGTTTGCTGA	
AACTAATCTG	CTATAGACAG	AGACTCCGGT	GAACCAATTT	TATTAGGATT	58200
TGATCAAATA	AACTCTCTCT	GACAAAGGAC	TGCTGAAAGA	GTAACATAAG	
GTTTGATGTT	TACTGAGTGC	ATAGTATGTG	CTAGATGCTG	GCCGTGGATG	58300
CCTCATAGAA	TCCTCCCAAC	AACTCATGAA	ATGACTACTG	TCATTAGGCC	
CAATACCCAG	ACGAGAAAGC	TGAGGGTAAG	ACAGGTTTCA	AGCTTGGCAG	58400
TCTGACTACA	GAGGCCACTG	GCTTAGCCCC	TGGGTTAGTC	TGCCTCTGTA	
GGATTGGGGG	CACGTAATTT	TGCTGTTTGG	GGTCTCATTT	GCCTTCTTAG	58500
AGATCACAAG	CCAAAGCTTT	TTATTCTAGA	GCCAAGGTCA	CGGAAGCCCA	
GAGGGCATCT	TGTGGCTCGG	GAGTAGCTCT	CTGCTGTCTT	CTCAGCTCTG	58600
CTGACAATAC	TTGAGATTTT	CAGATGTCAC	CAACCACCAA	GAGAGCTTGA	
TATGACTGTA	TATAGTATAG	TCATAAAGAA	CCTGAACCTG	ACCATATACT	58700
TATGTCAATG	GGAAAAATTT	TCATAGCTTC	AGATAGATTA	TATCTGGAGT	
GAAGAATCCT	GCCACCTATG	TATCTGGCAT	AGTGTGAGTC	CTCATAAATG	58800
CTTACTGGTT	TGAAGGGCAA	CAAAATAGTG	AACAGAGTGA	AAATCCCCAC	
TAAGATCCTG	GGTCCAGAAA	AAGATGGGAA	ACCTGTTTAG	CTCACCCGTG	58900
AGCCCATAGT	TAAAACTCTT	TAGACAACAG	GTTTTTTCCG	TTTACAGAGA	
ACAATAATAT	TGGGTGGTGA	GCATCTGTGT	GGGGGTGGG	GTGGGATAGG	59000
GGATACGGGG	AGAGTGGAGA	AAAAGGGGGC	ACAGGGTTAA	TGTGAAGTCC	
A					
AGGATCCCCC	TCTACATTTA	AAGTTGGTTT	AAGTTGGCTT	TAATTAATAG	59100
CAACTCTTAA	GATAATCAGA	ATTTTCTTAA	CCTTTTAGCC	TTACTGTGTA	
AAAGCCCTGT	GATCTTGTAC	AAATCATTTG	CTTCTTGGAT	AGTAATTTCT	59200
TTTACTAAAA	TGTGGGCTTT	TGACTAGATG	AATGTAAATG	TTCTTCTAGC	
TCTGATATCC	TTTATTCTTT	ATATTTTCTA	ACAGATTCTG	TGTAAGTGGGA	59300
TGAGCAGAGA	ACAAAAACAA	AATAATCCAG	TGAGAAAAGC	CCGTAAATAA	
ACTTTCAGAC	CAGAGATCTA	TTCTCTAGCT	TATTTTAAGC	TCAACTTAAA	59400
C T					
AGGAAGAACT	GTTCTCTGAT	TCTTTTCGCC	TTCAATACAC	TTAATGATTT	
A					
AACTCCACCC	TCCTTCAAAA	GAAACAGCAT	TTCCTACTTT	TATACTGTCT	59500
ATATGATTGA	TTTGCACAGC	TCATCTGGCC	AGAAGAGCTG	AGACATCCGT	
TCCCTACAA	GAAACTCTCC	CCGGTAAGTA	ACCTCTCAGC	TGCTTGGCCT	59600
GTTAGTTAGC	TTCTGAGATG	AGTAAAAGAC	TTTACAGGAA	ACCCATAGAA	
GACATTTGGC	AAACACCAAG	TGCTCATACA	ATTATCTTAA	AATATAATCT	59700
T					
TTAAGATAAG	GAAAGGGTCA	CAGTTTGGAA	TGAGTTTCAG	ACGGTTATAA	
CATCAAAGAT	ACAAAACATG	ATTGTGAGTG	AAAGACTTTA	AAGGGAGCAA	59800

FIGURE 1A

TAGTATTTTA	ATAACTAACA	ATCCTTACCT	CTCAAAGAA	AGATTTCAG	
C					
AGAGATGAGT	CTTAGCTGAA	ATCTTGAAAT	CTTATCTTCT	GCTAAGGAGA	59900
ACTAAACCCCT	CTCCAGTGAG	ATGCCCTTCTG	AATATGTGCC	CACAAGAAGT	
TGTGTCTAAG	TCTGGTTCTC	TTTTTCTTTT	TTCCTCCAGA	CAAGAGGGAA	60000
GCCTAAAAAT	GGTCAAAAT	AATATTAAAT	TACAAACGCC	AAATAAAAT	
TTCTCTAAT	ATATCAGTTT	CATGGCACAG	TTAGTATATA	ATTCTTTATG	60100
GTTCAAAAT	AAAAATGAGC	TTTTCTAGGG	GCTTCTCTCA	GCTGCCTAGT	
CTAAGGTGCA	GGGAGTTTGA	GACTCACAGG	GTTTAATAAG	AGAAAAATTCT	60200
CAGCTAGAGC	AGCTGAACTT	AAATAGACTA	GGCAAGACAG	CTGGTTATAA	
GACTAAACTA	CCCAGAATGC	ATGACATTCA	TCTGTGGTGG	CAGACGAAAC	60300
ATTTTTTATT	ATATTATTTT	TTGGGTATGT	ATGACAACCTC	TTAATTGTGG	
CAACTCAAAC	TACAAACACA	AACTTCACAG	AAAATGTGAG	GATTTTACAA	60400
TTGGCTGTTG	TCATCTATGA	CCTTCCCTGG	GACTTGGGCA	CCCGGCCATT	
TCACTCTGAC	TACATCATGT	CACCAAACAT	CTGATGGTCT	TGCCTTTAA	60500
TTCTCTTTTT	GAGGACTGAG	AGGGAGGGTA	GCATGGTAGT	TAAGAGTGCA	
GGCTTCCCGC	ATTCAAAATC	GGTTGCTTAC	TAGCTGTGTG	GCTTTGAGCA	60600
AGTTACTCAC	CCTCTCTGTG	CTTCAAGGTC	CTTGTCTGCA	AAATGTGAAA	
AATATTCCT	GCCTCATAAG	GTTGCCCTAA	GGATTAAATG	AATGAATGGG	60700
TATGATGCTT	AGAACAGTGA	TTGGCATCCA	GTATGTGCCC	TCGAGGCCCTC	
TTAATTATTA	CTGGCTTGCT	CATAGTGCAT	GTTCTTTGTG	GGCTAACTCT	60800
AGCGTCAATA	AAAATGTTAA	GACTGAGTTG	CAGCTGGGCA	TGGTGGCTCA	
TGCCTGTAA	CCCAGCATT	TAGGAGGCTG	AGGCAGGAGG	ATCGCTTGAG	60900
CCCAGGAGTT	CGAGACCAGC	CTGGGCAACA	TAGTGTGATC	TTGTATCTAT	
AAAAATAAAC	AAAATTAGCT	TGGTGTGGTG	GCGCCTGTAG	TCCCCAGCCA	61000
CTTGAGGGGG	TGAGGTGAGA	GGATTGCTTG	AGCCCCGGAT	GATCCAGGCT	
GCAGTGAGCC	ATGATCGTGC	CACCTGCACTC	CAGCCTGGGC	GACAGAGTGA	61100
GACCCGTCT	CACAACAACA	ACAGCAACAA	AAAGGCTGAG	CTGCACCATG	
CTTGACCCAG	TTTCTTAAAA	TTGTTGTCAA	AGCTTCATTC	ACTCCATGGT	61200
GCTATAGAGC	ACAAGATTTT	ATTTGGTGTG	ATGGTGCTTT	CATGAATTCC	
CCCAACAGAG	CCAAGCTCTC	CATCTAGTGG	ACAGGGAAGC	TAGCAGCAAA	61300
CCTTCCCTTC	ACTACAAAAC	TTCATTGCTT	GGCCAAAAAG	AGAGTTAATT	
CAATGTAGAC	ATCTATGTAG	GCAATTAAAA	ACCTATTGAT	GTATAAAACA	61400
T					
GTTTGCATTC	ATGGAGGGCA	ACTAAATACA	TTCTAGGACT	TTATAAAAGA	
TCACTTTTTA	TTTATGCACA	GGGTGGAACA	AGATGGATTA	TCAAGTGTCA	61500
[EXON 2: 61483..					
AGTCCAATCT	ATGACATCAA	TTATTATACA	TCGGAGCCCT	GCCAAAAAAT	
CAATGTGAAG	CAAATCGCAG	CCCGCCTCCT	GCCTCCGCTC	TACTCACTGG	61600
TGTTTCATCTT	TGGTTTTGTG	GGCAACATGC	TGGTCATCCT	CATCCTGATA	
A					
AACTGCAAAA	GGCTGAAGAG	CATGACTGAC	ATCTACCTGC	TCAACCTGGC	61700
CATCTCTGAC	CTGTTTTTCC	TTCTTACTGT	CCCCTTCTGG	GCTCACTATG	
C					
CTGCCGCCCA	GTGGGACTTT	GGAAATACAA	TGTGTCAACT	CTTGACAGGG	61800
CTCTATTTTA	TAGGCTTCTT	CTCTGGAATC	TTCTTCATCA	TCCTCCTGAC	
T					
AATCGATAGG	TACCTGGCTG	TCGTCCATGC	TGTGTTTGCT	TTAAAAGCCA	61900
GGACGGTCAC	CTTTGGGGTG	GTGACAAGTG	TGATCACTTG	GGTGGTGGCT	
GTGTTTGCGT	CTCTCCCAGG	AATCATCTTT	ACCAGATCTC	AAAAAGAAGG	62000
TCTTCATTAC	ACCTGCAGCT	CTCATTTTCC	ATACAGTCAG	TATCAATTCT	
G					
GGAAGAATTT	CCAGACATTA	AAGATAGTCA	TCTTGGGGCT	GGTCCTGCCG	62100

FIGURE 1B

CTGCTTGTCA	TGGTCATCTG	CTACTCGGGA	ATCCTAAAAA	CTCTGCTTCG	
				A	
GTGTGCAAAAT	GAGAAGAAGA	GGCACAGGGC	TGTGAGGCTT	ATCTTCACCA	62200
TCATGATTGT	TTATTTTCTC	TTCTGGGCTC	CCTACAACAT	TGTCCTTCTC	
CTGAACACCT	TCCAGGAATT	CTTTGGCCTG	AATAATTGCA	GTAGCTCTAA	62300
CAGGTTGGAC	CAAGCTATGC	AGGTGACAGA	GACTCTTGGG	ATGACGCACT	
GCTGCATCAA	CCCCATCATC	TATGCCCTTG	TCGGGGAGAA	GTTTCAGAAAC	62400
TACCTCTTAG	TCTTCTTCCA	AAAGCACATT	GCCAAACGCT	TCTGCAAATG	
CTGTTCTATT	TTCCAGCAAG	AGGCTCCCGA	GCGAGCAAGC	TCAGTTTACA	62500
CCCATCCAC	TGGGGAGCAG	GAAATATCTG	TGGGCTTGTG	ACACGGACTC	
	..62541]				
AAGTGGGCTG	GTGACCCAGT	CAGAGTTGTG	CACATGGCTT	AGTTTTTCATA	62600
CACAGCCTGG	GCTGGGGGTG	GGGTGGGAGA	GGTCTTTTTT	AAAAGGAAGT	
TACTGTTATA	GAGGGTCTAA	GATTCATCCA	TTTATTGCGC	ATCTGTTTAA	62700
AGTAGATTAG	ATCTTTTAAG	CCCATCAATT	ATAGAAAGCC	AAATCAAAAT	
ATGTTGATGA	AAAATAGCAA	CCTTTTATC	TCCCCTTCAC	ATGCATCAAG	62800
TTATTGACAA	ACTCTCCCTT	CACTCCGAAA	GTTCCCTATG	TATATTTAAA	
AGAAAGCCTC	AGAGAATTGC	TGATTCTTGA	GTTTAGTGAT	CTGAACAGAA	62900
ATACCAAAAT	TATTTTCAGAA	ATGTACAACT	TTTTACCTAG	TACAAGGCAA	
CATATAGGTT	GTAAATGTGT	TTAAAACAGG	TCTTTGTCTT	GCTATGGGGA	63000
GAAAAGACAT	GAATATGATT	AGTAAAGAAA	TGACACTTTT	CATGTGTGAT	
TTCCCCTCCA	AGGTATGGTT	AATAAGTTTC	ACTGACTTAG	AACCAGGCGA	63100
GAGACTTGTT	GCCTGGGAGA	GCTGGGGGAG	CTTCTTAAAT	GAGAAGGAAT	
TTGAGTTGGA	TCATCTATTG	CTGGCAAAGA	CAGAAGCCTC	ACTGCAAGCA	63200
CTGCATGGGC	AAGCTTGGCT	GTAGAAGGAG	ACAGAGCTGG	TTGGGAAGAC	
ATGGGGAGGA	AGGACAAGGC	TAGATCATGA	AGAACCCTGA	CGGCATTGCT	63300
CCGTCTAAGT	CATGAGCTGA	GCAGGGAGAT	CCTGGTTGGT	GTTGCAGAAG	
GTTTACTCTG	TGGCCAAAGG	AGGGTCAGGA	AGGATGAGCA	TTTAGGGCAA	63400
GGAGACCACC	AACAGCCCTC	AGGTCAGGGT	GAGGATGGCC	TCTGCTAAGC	
TCAAGGCGTG	AGAGGGGAA	GGAGGGAGGT	ATTCGTAAGG	ATGGGAAGGA	63500
GGGAGGTATT	CGTGCAAGCAT	ATGAGGATGC	AGAGTCAGCA	GAAGTGGGGT	
GGATTTGGTT	TGGAAGTGAG	GGTCAGAGAG	GAGTCAGAGA	GAATCCCTAG	63600
TCTTCAAGCA	GATTGGAGAA	ACCCTTGAAA	AGACATCAAG	CACAGAAGGA	
GGAGGAGGAG	GTTTAGGTCA	AGAAGAAGAT	GGATTGGTGT	AAAAGGATGG	63700
GTCTGGTTTG	CAGAGCTTGA	ACACAGTCTC	ACCCAGACTC	CAGGCTGTCT	
TTCCTGAAT	GCTTCTGACT	TCATAGATTT	CCTTCCCATC	CCAGCTGAAA	63800
TACTGAGGGG	TCTCCAGGAG	GAGACTAGAT	TTATGAATAC	ACGAGGTATG	
AGGTCTAGGA	ACATACTTCA	GCTCACACAT	GAGATCTAGG	TGAGGATTGA	63900
TTACCTAGTA	GTCATTTTCA	GGGTTGTTGG	GAGGATTCTA	TGAGGCAACC	
ACAGGCAGCA	TTTAGCACAT	ACTACACATT	CAATAAGCAT	CAAACCTCTTA	64000
GTTACTCATT	CAGGGATAGC	ACTGAGCAAA	GCATTGAGCA	AAGGGGTCCC	
ATAGAGGTGA	GGGAAGCCTG	AAAAACTAAG	ATGCTGCCTG	CCCAGTGCAC	64100
ACAAGTGTAG	GTATCATTTT	CTGCATTTAA	CCGTCAATAG	GCAAAGGGGG	
GAAGGGACAT	ATTCAATTTG	AAATAAGCTG	CCTTGAGCCT	TAAAACCCAC	64200
AAAAGTACAA	TTTACCAGCC	TCCGTATTTT	AGACTGAATG	GGGGTGGGGG	
GGGCGCCTTA	GGTACTTATT	CCAGATGCCT	TCTCCAGACA	AACCAGAAGC	64300
AACAGAAAAA	ATCGTCTCTC	CCTCCCTTTG	AAATGAATAT	ACCCCTTAGT	
GTTTGGGTAT	ATTCAATTTCA	AAGGGAGAGA	GAGAGGTTTT	TTTCTGTTCT	64400
TTCTCATATG	ATTGTGCACA	TACTTGAGAC	TGTTTTGAAT	TTGGGGGATG	
GCTAAAACCA	TCATAGTACA	GGTAAGGTGA	GGGAATAGTA	AGTGGTGAGA	64500
ACTACTCAGG	GAATGAAGGT	GTCAGAATAA	TAAGAGGTGC	TACTGACTTT	
CTCAGCCTCT	GAATATGAAC	GGTGAGCATT	GTGGCTGTCA	GCAGGAAGCA	64600
ACGAAGGGAA	ATGTCTTTCC	TTTTGCTCTT	AAGTTGTGGA	GAGTGCAACA	
GTAGCATAGG	ACCCTACCCT	CTGGGCCAAG	TCAAAGACAT	TCTGACATCT	64700

FIGURE 1C

4/6

TAGTATTTGC	ATATTCTTAT	GTATGTGAAA	GTTACAAATT	GCTTGAAAGA	
AAATATGCAT	CTAATAAAAA	ACACCTTCTA	AAATAATTCA	TTATATTCTT	64800
GCTCTTTCAG	TCAAGTGTAC	ATTTAGAGAA	TAGCACATAA	AACTGCCAGA	
GCATTTTATA	AGCAGCTGTT	TTCTTCCTTA	GTGTGTGTGC	ATGTGTGTGT	64900
GATGTATACA	AAGAGAGAGA	TAATTGTATT	TTTGTATTTT	CTTTTAAATA	
ATTTTTTAAA	TTGACCCCTT	TCCTGAGACA	AATTGCCAGA	ATAGTTTGTA	65000
TTTAGAGATG	GTACCTCTAA	GAGTAAGGTT	GCTGGTTGCT	GAGCAATTGA	
CTTGAAACT	TTTAAATTC	AAATTTAAT	TCCACTACTC	AAAAGAATTG	65100
CCATGTTTTA	AAAAAGAGAA	TTGGTGCCAT	AAGTTAGTTG	TCTATGTTTG	
AAAATGAAGA	AGATATGCAA	CGTCATGGCC	TGGTCACCTA	CCCGCAGCCC	65200
TGAGTTGTAG	GCACATCATA	TGTGAGAATG	AGGATGCTTT	TCTTTCATTT	
AAAATCCCTC	CCCAAAACTT	GGCTCTAATT	GCAGTCATGA	CAATCATGTA	65300
CATTTGGATT	TATGTGCACG	AGTCTCTTAC	CCTGAGAGAG	GACAGGTGCT	
ACAGGTGGAG	GGGACCCGTC	TGGGTCACGT	TCACATTTTG	AACATGCTGG	65400
TTTTTCAGTCA	CTGCACACTC	ATCTCCCAGC	GCAGGTCATG	GGCAGCAGAT	
GCAAAAGCTG	CCCGTGGTCC	TATTTGGAGG	TGCATGAAAT	GAGCAGAAGA	65500
CAGAATAGCT	TGATCTGACT	AGAAGGGCAG	CTTGTCCTTA	CCAAGACTTG	
AAGGATTGCC	TTTCATCTGT	TAGGGTAAAA	GGTAGAATGA	ACCAAGGAAG	65600
GGCAGGAGGG	GGCTGGGGTT	AGGGTAGAAG	GAAGGGGCCA	TGGAGAAGGG	
AGATCCATCC	CATAGGAGGA	AGGCAGTGCG	GCAGGGAGGT	TTGAAGGTAT	65700
CAGCTTTTGT	GGCTGACATA	CATGCAGTCA	TGTCAATTGC	TCGTTTTTCC	
TTTTCCATCT	TATTAAATGT	CTTCCAATGT	TAGCACGAAG	AAAAGCTATT	65800
TGCAGTGTG	CCAGCCTTTC	CAGAGCCCGT	CCCCATTACC	TCCCCAGGCC	
CATGCCTTTA	CTCCTTGGAG	TTTCAACTCA	CGACCTTCAG	GATCTGACTT	65900
TATTCACCAA	CTCTGGGGTG	AACGTACCTT	CTGTCTCCAC	CCAGAGGTCT	
CTATCAAAGA	GGAGATTGCA	TGCCATGGAT	AAAGTCAAAG	TAGAGGTGAC	66000
TGTCCTTAGG	AAGAGTAATG	TGAAAATTCA	TAAACTGGGA	TTCTGTTTAC	
ATTTTGTACT	CCAGGGGTTC	TTAGTTTAAA	TCGCTCTGAA	TAAATTAAGA	66100
TGCAATGGCA	TTTCAACTGT	TATGATTAAA	TTTACAAATC	ATTTATTTTC	
TATCACGGGG	AGAGATAGAG	CTCCAAATGC	AAACATAACT	GCTCAAGTGT	66200
TAACACTTAT	AATGAAAACA	TAAGAATTAC	CACCAACTAC	CCTGGGGGCT	
AGAAGCAGAA	ATGTGAACCA	GAAAACAAAT	CATGAACTTT	CCTTTTTTTT	66300
TTTGAGATGG	AGTCTCGCTC	TGTTGCCCG	GCTGGAGTGC	AATGGTGCGA	
TCTCGGCTCA	CTGCAACCAC	TGCCTCCCGG	GTTCAAGCAA	TTCTCCTGCC	66400
TCAGCCTCCT	GAGTAGCTGG	GACTIONAGGC	ATGCACCACC	ACGCCTGGGT	
AATTTTTTGT	ATTTTGTAGTA	GAGACAGGGT	TTCACCGTAT	TAGCCAGGAT	66500
GCTCTCGATC	TCCTGACCTC	GTGATCTGCC	CGCCTCGGCC	T	66541

FIGURE 1D

5/6

POLYMORPHISMS IN THE CODING SEQUENCE OF CCR5

ATGGATTATC	AAGTGTCAAG	TCCAATCTAT	GACATCAATT	ATTATACATC	
GGAGCCCTGC	CAAAAAATCA	ATGTGAAGCA	AATCGCAGCC	CGCCTCCTGC	100
CTCCGCTCTA	CTCACTGGTG	TTCATCTTTG	GTTTTGTGGG	CAACATGCTG	
GTCATCCTCA	TCCTGATAAA	CTGCAAAAGG	CTGAAGAGCA	TGACTGACAT	200
	A				
CTACCTGCTC	AACCTGGCCA	TCTCTGACCT	GTTTTTCCTT	CTTACTGTCC	
		C			
CCTTCTGGGC	TCACTATGCT	GCCGCCAGT	GGGACTTTGG	AAATACAATG	300
TGTCAACTCT	TGACAGGGCT	CTATTTTATA	GGCTTCTTCT	CTGGAATCTT	
		T			
CTTCATCATC	CTCCTGACAA	TGCATAGGTA	CCTGGCTGTC	GTCCATGCTG	400
TGTTTGCTTT	AAAAGCCAGG	ACGGTCACCT	TTGGGGTGGT	GACAAGTGTG	
ATCACTTGGG	TGGTGGCTGT	GTTTGCGTCT	CTCCCAGGAA	TCATCTTTAC	500
CAGATCTCAA	AAAGAAGGTC	TTCATTACAC	CTGCAGCTCT	CATTTTCCAT	
			G		
ACAGTCAGTA	TCAATTCTGG	AAGAATTTCC	AGACATTAAA	GATAGTCATC	600
TTGGGGCTGG	TCCTGCCGCT	GCTTGTCATG	GTCATCTGCT	ACTCGGGAAT	
CCTAAAAACT	CTGCTTCGGT	GTCGAAATGA	GAAGAAGAGG	CACAGGGCTG	700
	A				
TGAGGCTTAT	CTTCACCATC	ATGATTGTTT	ATTTTCTCTT	CTGGGCTCCC	
TACAACATTG	TCCTTCTCCT	GAACACCTTC	CAGGAATTCT	TTGGCCTGAA	800
TAATTGCAGT	AGCTCTAACA	GGTTGGACCA	AGCTATGCAG	GTGACAGAGA	
CTCTTGGGAT	GACGCACTGC	TGCATCAACC	CCATCATCTA	TGCCTTTGTC	900
GGGGAGAAGT	TCAGAAACTA	CCTCTTAGTC	TTCTTCCAAA	AGCACATTGC	
CAAACGCTTC	TGCAAATGCT	GTTCTATTTT	CCAGCAAGAG	GCTCCCGAGC	1000
GAGCAAGCTC	AGTTTACACC	CGATCCACTG	GGGAGCAGGA	AATATCTGTG	
GGCTTGTGA					1059

FIGURE 2

6/6

ISOFORMS OF THE CCR5 PROTEIN

MDYQVSSPIY	DINYYTSEPC	QKINVKQIAA	RLLPPLYSLV	FIFGFVGNML	
VILILINCKR	LKSMTDIYLL	NLAISDLFFL	LTVPFWAHYA	AAQWDFGNTM	100
Q					
CQLLTGLYFI	GFFSGIFFII	LLTIDRYLAV	VHAVFALKAR	TVTFGVVTSV	
ITWVVAVFAS	LPGIIFTRSQ	KEGLHYTCSS	HFPYSQYQFW	KNFQTLKIVI	200
L					
LGLVLPLLVM	VICYSGILKT	LLRCRNEKKR	HRAVRLIFTI	MIVYFLFWAP	
Q					
YNIVLLLNTF	QEFFGLNNCS	SSNRLDQAMQ	VTETLGMTHC	CINPIIYAFV	300
GEKFRNYLLV	FFQKHIAKRF	CKCCSIFQQE	APERASSVYT	RSTGEQEISV	
GL					352

FIGURE 3

SEQUENCE LISTING

<110> Genaissance Pharmaceuticals Inc.

Choi, Julie

Kliem, Stefanie E.

Koshy, Beena

<120> Haplotypes of the CCR5 Gene

<130> MWH-0505CT CCR5

<140> TBA

<141> 2001-04-04

<150> 60/194,361

<151> 2000-04-05

<160> 24

<170> PatentIn Ver: 2.1

<210> 1

<211> 9141

<212> DNA

<213> Homo sapien

<400> 1

```

gtatgtcttc acatcctcca atacccta atacccta gacagacaaa cagaacatgg caaagcctca 60
gctctgcatg gtgaaagtaa gaaccagcaa ttgccacaaa cagaaatata gtgttggtcc 120
ggcagcctcc gggggttctg cacaagtga ttaccagtga atacaaggct atctatcttt 180
cgaaaaacca acgttgtatt tatgtctatt tctataaaat tttatattaa tttatttgtt 240
acctatcttt gaactctttc aaaagcacac tttatatttc cctgcttaaa cagtcccccg 300
agggtgggtg cccaaaaggc tctacacttg ttatcattcc ctctccacca caggcatatt 360
gagtaagttt gtatttgggt ttttttaaaa cctccactct acagttaaga aaactaaggc 420
acagagcttc aataatttgg tcagagccaa gtacagtaa tgaagctgga ggttaaacc 480
agcagcatga ctgcagttct taatcaatgc ctttgaatt gcacatatgg gatgaactag 540
aacattttct cgatgattcg ctgtccttgt tatgattatg ttactgagct ctgttgtagc 600
acagacatat gtccctatat ggggcggggt tgggggtgtc ttgatcgctg ggctatttct 660
atactgttct ggcttttccc aagcagtcac ttctttctat tctccaagca ccagcaatta 720
gctttacctt ttcagcttct agtttgctga aactaatctg ctatagacag agactccggt 780
gaaccaattt tattaggatt tgatcaaata aactctctct gacaaaggac tgctgaaaga 840
gtaactaaga gtttgatgtt tactgagtgc atagtatgtg ctatagctgt gccgtggatg 900
cctcatagaa tcctcccaac aactcatgaa atgactactg tcattcagcc caataccag 960
acgagaaaagc tgagggttaag acaggtttca agcttggcag tctgactaca gaggccaotg 1020
gcttagcccc tgggttagtc tgcctctgta ggattggggg cacgtaattt tgctgttttg 1080
ggtctcattt gccttcttag agatcacaag ccaaagcttt ttattctaga gccaagggtc 1140
cggaagccca gagggcatct tgtggctcgg gagtagctct ctgctgtctt ctcagctctg 1200
ctgacaatac ttgagatttt cagatgtcac caaccaccaa gagagcttga tatgactgta 1260

```

tatagtatatag tcataaagaa cctgaacttg accatatact tatgtcatgt ggaaaatttc 1320
 tcatagcttc agatagatta tatctggagt gaagaatcct gccacctatg tatctggcat 1380
 agtgtgagtc ctcataaatg cttactgggt tgaagggcaa caaaatagtg aacagagtga 1440
 aaatccccac taagatcctg ggtccagaaa aagatgggaa acctgttttag ctcacccgtg 1500
 agcccatagt taaaactctt tagacaacag gttttttccg ttacagaga acaataatat 1560
 tgggtggtga gcatctgtgt gggggttggg gtgggatagg ggatacgggg agagtggaga 1620
 aaaagggggc acaggggttaa tgtgaagtcc aggatcccc tctacattta aagttggttt 1680
 aagttggctt taattaatag caactcttaa gataatcaga attttcttaa ccttttagcc 1740
 ttactgttga aaagccctgt gatcttgtac aaatcatttg cttcttggat agtaatttct 1800
 ttactaaaa tgtgggcttt tgactagatg aatgtaaatg ttcttctagc tctgatatcc 1860
 ttattctttt atattttcta acagattctg tgtagtggga tgagcagaga acaaaaacaa 1920
 aataatccag tgagaaaagc ccgtaaataa actttcagac cagagatcta ttctctagct 1980
 tattttaagc tcaacttaaa aggaagaact gttctctgat tcttttcgcc ttcaatacac 2040
 ttaatgattt aactccaccc tcttcaaaa gaaacagcat ttctacttt tatactgtct 2100
 atatgattga ttgacagc tcatctggcc agaagagctg agacatccgt tccccacaa 2160
 gaaactctcc ccggtaaagta acctctcagc tgcttggcct gttagttagc ttctgagatg 2220
 agtaaaagac ttacagga acccatagaa gacatttggc aaacaccaag tgctcataca 2280
 attatcttaa aatataatct ttaagataag gaaagggca cagtttgga tgagtttcag 2340
 acggttataa catcaaagat acaaaacatg attgtgagt aaagacttta aaggagcaa 2400
 tagtatttta ataactaaca atccttacct ctcaaaagaa agatttgag agagatgagt 2460
 cttagctgaa atcttgaaat cttatcttct gctaaggaga actaaacct ctccagttag 2520
 atgcctctg aatatgtgcc cacaagaagt tgtgtctaag tctggttctc tttttcttt 2580
 ttctccaga caagagggaa gcctaaaaat ggtcaaaatt aatattaaat tacaacgcc 2640
 aaataaaatt ttctctaata atatcagtt catggcacag ttagtatata attctttatg 2700
 gttcaaaatt aaaaatgagc ttttctagg gcttctctca gctgcctagt ctaagggtgca 2760
 gggagttaga gactcacagg gtttaataag agaaaattct cagctagagc agctgaactt 2820
 aaatagacta ggcaagacag ctggttataa gactaaacta ccagaatgc atgacattca 2880
 tctgtggtg cagacgaaac attttttatt atattatttc ttgggtatgt atgacaactc 2940
 ttaattgtg caactcaaac tacaacaca aacttcacag aaaatgtgag gattttacaa 3000
 ttggctgtg tcatctatga ccttccctgg gacttgggca cccggccatt tcaactctgac 3060
 tacatcatgt caccaacat ctgatggtct tgccttttaa ttctctttt gaggactgag 3120
 agggagggtg gcatggtagt taagagtga ggcttccgc attcaaaatc ggttgcttac 3180
 tagctgtgtg gctttgagca agttactcac cctctctgtg cttcaaggtc cttgtctgca 3240
 aaatgtgaaa aatatttctt gcctcataag gttgcctaa ggattaaatg aatgaatggg 3300
 tatgatgctt agaacagtga ttggcatcca gtatgtgcc tcgaggcctc ttaattatta 3360
 ctggcttct catagtgc atgttcttgt ggctaactct agcgtcaata aaaatgttaa 3420
 gactgagttg cagctgggca tgggtggctca tgcctgtaat ccagcattc taggaggctg 3480
 aggcaggagg atcgttgag cccaggagtt cgagaccagc ctgggcaaca tagtgtgatc 3540
 ttgtatctat aaaaataaac aaaattagct tgggtggtg gcgctgtag tccccagcca 3600
 cttggagggg tgaggtgaga ggattgcttg agcccggtat gatccaggct gcagtgagcc 3660
 atgatcgtgc cactgcactc cagcctgggc gacagagtga gaccctgtct cacaacaaca 3720
 acagcaacaa aaaggctgag ctgcacatg cttgaccag tttcttaaaa ttgttgtcaa 3780
 agcttcattc actccatggt gctatagagc acaagatttt atttggtgag atggtgcttt 3840
 catgaattcc cccaacagag ccaagctctc catctagtgg acagggaagc tagcagcaa 3900
 ccttccctc actacaaaac ttcatgtctt ggccaaaaag agagttaatt caatgtagac 3960
 atctatgtag gcaattaaaa acctattgat gtataaaca gtttgcatc atggagggca 4020
 actaaataca ttctaggact ttataaaga tcacttttta ttatgcaca gggtggaaca 4080
 agatggatta tcaagtgtca agtccaatct atgacatcaa ttattatata tggagccct 4140

gccaaaaaat caatgtgaag caaatcgag cccgcctcct gcctccgctc tactcactgg 4200
 tgttcatctt tggttttgtg ggcaacatgc tggatcatct catcctgata aactgcaaaa 4260
 ggctgaagag catgactgac atctacctgc tcaacctggc catctctgac ctgtttttcc 4320
 ttcttactgt ccccttctgg gctcactatg ctgccgccca gtgggacttt ggaaatacaa 4380
 tgtgtcaact cttagacagg ctctatttta taggcttctt ctctggaatc ttcttcatca 4440
 tcctcctgac aatcgatagg tacctggctg tgcctcatgc tgtgtttgct ttaaaagcca 4500
 ggacggtcac ctttgggtg gtgacaagt tgatcacttg ggtgggtggc gtgtttgcgt 4560
 ctctcccagg aatcatcttt accagatctc aaaaagaagg tcttcattac acctgcagct 4620
 ctcatcttcc atacagtcag tatcaattct ggaagaattt ccagacatta aagatagtca 4680
 tcttggggct ggtcctggcg ctgcttgta tggatcatct ctactcggga atcctaataa 4740
 ctctgcttcg gtgcgaaat gagaagaaga ggcacagggc tgtgaggctt atcttcacca 4800
 tcatgattgt ttattttctc ttctgggctc cctacaacat tgctcttctc ctgaacacct 4860
 tccaggaatt ctttggcctg aataattgca gtagctctaa cagggttgac caagctatgc 4920
 aggtgacaga gactcttggg atgacgcact gctgcatcaa ccccatcatc tatgcctttg 4980
 tcggggagaa gttcagaaac tacctcttag tcttcttcca aaagcacatt gccaaacgt 5040
 tctgcaaatg ctgttctatt ttccagcaag aggtctccga gcgagcaagc tcagtttaca 5100
 cccgatccac tggggagcag gaaatatctg tgggcttggt acacggactc aagtgggctg 5160
 gtgacccagt cagagtgtg cacatggctt agttttcata cacagcctgg gctgggggtg 5220
 ggggtggaga ggtctttttt aaaaggaagt tactgttata gaggttetaa gattcatcca 5280
 tttatttggc atctgtttaa agtagattag atcttttaag cccatcaatt atagaaagcc 5340
 aaatcaaat atgttgatga aaaatagcaa cctttttatc tccccttcac atgcatcaag 5400
 ttattgacaa actctccctt cactccgaaa gttccttatg tatattttaa agaaagcctc 5460
 agagaattgc tgattcttga gtttagtgat ctgaacagaa ataccaaaat tatttcagaa 5520
 atgtacaact ttttacctag tacaaggcaa catatagggt gtaaatgtgt ttaaaacagg 5580
 tctttgtctt gctatgggga gaaaagacat gaatatgatt agtaaagaaa tgacactttt 5640
 catgtgtgat ttcccctcca aggtatggtt aataagtttc actgacttag aaccaggcga 5700
 gagacttgtg gcctgggaga gctggggaag cttctttaat gagaaggaat ttgagttgga 5760
 tcatctattg ctggcaaga cagaagcctc actgcaagca ctgcatgggc aagcttggct 5820
 gtagaaggag acagagctgg ttgggaagac atggggagga aggacaaggc tagatcatga 5880
 agaacttga cggcattgct ccgtctaagt catgagctga gcagggagat cctggttgg 5940
 gttgcagaag gtttactctg tggccaaagg agggctagga aggatgagca tttagggcaa 6000
 ggagaccacc aacagccctc aggtcagggt gaggatggcc tctgctaagc tcaaggcgtg 6060
 aggatgggaa ggaggagggt attcgtaagg atgggaagga gggagggtatt cgtgcagcat 6120
 atgaggatgc agagtcagca gaactgggtt ggatttgggt tggaaagtga ggtcagagag 6180
 gagtcagaga gaatccctag tcttcaagca gattggagaa acccttgaaa agacatcaag 6240
 cacagaagga ggaggaggag gtttaggtca agaagaagat ggattggtgt aaaaggatgg 6300
 gtctggtttg cagagcttga acacagctc acccagaçtc caggctgtct ttcactgaat 6360
 gcttctgact tcatagattt ccttcccac ccagctgaaa tactgagggg tctccaggag 6420
 gagactagat ttatgaatac acgaggatg aggtctagga acatacttca gctcacacat 6480
 gagatctagg tgaggattga ttacctagta gtcatttcat gggttggttg gaggattcta 6540
 tgaggcaacc acaggcagca tttagcacat actacacatt caataagcat caaactctta 6600
 gttactcatt cagggatagc actgagcaaa gcaattgagca aaggggtccc atagaggta 6660
 gggaagcctg aaaaactaag atgctgcctg ccagtgac acaagtgtag gtatcatctt 6720
 ctgcatttaa ccgtcaatag gcaaaggggg gaaggacat attcatttgg aaataagctg 6780
 ccttgagcct taaaaccac aaaagtacaa tttaccagcc tccgtatttc agactgaatg 6840
 ggggtggggg gggcgctta ggtacttatt ccagatgcct tctccagaca aaccagaagc 6900
 aacagaaaa atcgtctctc cctccctttg aaatgaatat aecccttagt gtttgggtat 6960
 attcatttca aaggagaga gagaggtttt ttctgttct ttctcatatg attgtgcaca 7020

tacttgagac tgttttgaat ttgggggatg gctaaaacca tcatagtaca ggtaaggtga 7080
 gggaatagta agtggtgaga actactcagg gaatgaaggt gtcagaataa taagaggtgc 7140
 tactgacttt ctcagcctct gaatatgaac ggtgagcatt gtggctgtca gcaggaagca 7200
 acgaagggaa atgtctttcc ttttgctctt aagttgtgga gagtgaaca gtagcatagg 7260
 accctaccct ctgggccaag tcaaagacat tctgacatct tagtatttgc atattcttat 7320
 gtatgtgaaa gttacaaatt gcttgaaaga aaatatgcat ctaataaaaa acaccttcta 7380
 aaataattca ttatattctt gctctttcag tcaagtgtac atttagagaa tagcacataa 7440
 aactgccaga gcattttata agcagctgtt ttcttcctta gtgtgtgtgc atgtgtgtgt 7500
 gatgtataca aagagagaga taattgtatt tttgtatttt cttttaaata atttttaaaa 7560
 ttgacccttt tcctgagaca aattgccaga atagtttcta tttagagatg gtacctctaa 7620
 gagtaagggt gctggttgc gagcaattga cttgaaaact tttaaaattc aaattttaat 7680
 tccactactc aaaagaattg ccatgtttta aaaaagagaa ttggtgccat aagttagttg 7740
 tctatgtttg aaaaatgaaga agatatgcaa cgtcatggcc tggtcactta cccgcagccc 7800
 tgagttgtag gcacatcata tgtgagaatg aggatgcttt tctttcattt aaaatccctc 7860
 cccaaaactt ggctctaatt gcagtcata caatcatgta catttggtt tatgtgcacg 7920
 agtctcttac cctgagagag gacaggtgct acaggtggag gggacccgct tgggtcacgt 7980
 tcacattttg aacatgctgg ttttcagtca ctgcacactc atctcccagc gcaggatcatg 8040
 ggcagcagat gcaaaagctg cccgtggtcc tatttggagg tgcataaaat gagcagaaga 8100
 cagaatagct tgatctgact agaagggcag cttgtcccta ccaagacttg aaggattgcc 8160
 tttcatctgt tagggtaaaa ggtagaatga accaaggaag ggcaggaggg ggctgggggt 8220
 agggtagaag gaaggggcca tggagaagg agatccatcc cataggagga aggcagtgcg 8280
 gcaggagggt ttgaaggat cagcttttgt ggctgacata catgcagtca tgtcaattgc 8340
 tcgtttttcc ttttccatct tattaatgt cttccaatgt tagcacgaag aaagactatt 8400
 tgcagtgttg ccagcctttc cagagcccgt cccattacc tcccaggcc catgccttta 8460
 ctcttggag tttcaactca cgaccttcag gatctgactt tattcaccaa ctctgggggtg 8520
 aacgtacctt ctgtctccac ccagaggtct ctatcaaaga ggagattgca tgccatggat 8580
 aaagtcaaag tagaggtgac tgtccttagg aagagtaatg tgaaaattca taaactggga 8640
 ttctgtttac attttgtact ccaggggttc ttagtttaaa tgcctctgaa taaattaaga 8700
 tgcaatggca tttcaactgt tatgattaaa ttacaaaatc atttattttc tatcacgggg 8760
 agagatagag ctccaaatgc aaacataact gctcaagtgt taacacttat aatgaaaaca 8820
 taagaattac caccaactac cctgggggct agaagcagaa atgtgaacca gaaaacaaat 8880
 catgaacttt cttttttttt tttgagatgg agtctcgctc tgttgccag gctggagtgc 8940
 aatggtgcga tctcggtca ctgcaaccac tgccctccgg gttcaagcaa ttctcctgcc 9000
 tcagcctcct gagtagctgg gactacaggg atgcaccacc acgcctgggt aattttttgt 9060
 attttttagta gagacagggt ttcaccgtat tagccaggat gctctcgatc tctgacctc 9120
 gtgatctgcc cgcctcggc t 9141

<210> 2

<211> 1059

<212> DNA

<213> Homo sapien

<400> 2

atggattatc aagtgtcaag tccaatctat gacatcaatt attatacatc ggagccctgc 60
 caaaaaatca atgtgaagca aatcgagcc cgcctcctgc ctccgctcta ctactgggtg 120
 ttcattcttg gttttgtggg caacatgctg gtcacctca tcctgataaa ctgcaaaagg 180
 ctgaagagca tgactgacat ctacctgctc aacctggcca tctctgacct gtttttcctt 240

ctactgtcc cttctgggc tcaactgct gccgccagt gggactttgg aaatacaatg 300
 tgtcaactct tgacagggt ctattttata ggcttcttct ctggaatctt cttcatcatc 360
 ctcccgacaa tcgataggta cctggctgtc gtccatgctg tgtttgcttt aaaagccagg 420
 acggtcacct ttgggggtgt gacaagtgtg atcacttggg tggaggctgt gtttgctct 480
 ctcccaggaa tcatctttac cagatctcaa aaagaaggtc ttcattacac ctgcagctct 540
 cattttccat acagtcaagta tcaattctgg aagaatttcc agacattaaa gatagtcac 600
 ttggggctgg tcctgccgct gcttgtcatg gtcactgct actcggaat cctaaaaact 660
 ctgcttcggt gtcgaaatga gaagaaggag cacagggtg tgaggcttat cttcaccatc 720
 atgattgttt attttctctt ctgggctccc tacaacattg tccttctcct gaacaccttc 780
 caggaattct ttggcctgaa taattgcagt agctctaaca ggttgacca agctatgcag 840
 gtgacagaga ctcttgggat gacgactgc tgcacaaacc ccatcatcta tgcctttgtc 900
 ggggagaagt tcagaaacta cctcttagtc ttcttccaaa agcaoattgc caaacgcttc 960
 tgcaaatgct gttctatttt ccagcaagag gctcccgagc gagcaagctc agtttacacc 1020
 cgatccactg gggagcagga aatatctgtg ggcttgtga 1059

<210> 3

<211> 352

<212> PRT

<213> Homo sapien

<400> 3

Met Asp Tyr Gln Val Ser Ser Pro Ile Tyr Asp Ile Asn Tyr Tyr Thr
 1 5 10 15

Ser Glu Pro Cys Gln Lys Ile Asn Val Lys Gln Ile Ala Ala Arg Leu
 20 25 30

Leu Pro Pro Leu Tyr Ser Leu Val Phe Ile Phe Gly Phe Val Gly Asn
 35 40 45

Met Leu Val Ile Leu Ile Leu Ile Asn Cys Lys Arg Leu Lys Ser Met
 50 55 60

Thr Asp Ile Tyr Leu Leu Asn Leu Ala Ile Ser Asp Leu Phe Phe Leu
 65 70 75 80

Leu Thr Val Pro Phe Trp Ala His Tyr Ala Ala Ala Gln Trp Asp Phe
 85 90 95

Gly Asn Thr Met Cys Gln Leu Leu Thr Gly Leu Tyr Phe Ile Gly Phe
 100 105 110

Phe Ser Gly Ile Phe Phe Ile Ile Leu Leu Thr Ile Asp Arg Tyr Leu
 115 120 125

Ala Val Val His Ala Val Phe Ala Leu Lys Ala Arg Thr Val Thr Phe
 130 135 140

Gly Val Val Thr Ser Val Ile Thr Trp Val Val Ala Val Phe Ala Ser
 145 150 155 160
 Leu Pro Gly Ile Ile Phe Thr Arg Ser Gln Lys Glu Gly Leu His Tyr
 165 170 175
 Thr Cys Ser Ser His Phe Pro Tyr Ser Gln Tyr Gln Phe Trp Lys Asn
 180 185 190
 Phe Gln Thr Leu Lys Ile Val Ile Leu Gly Leu Val Leu Pro Leu Leu
 195 200 205
 Val Met Val Ile Cys Tyr Ser Gly Ile Leu Lys Thr Leu Leu Arg Cys
 210 215 220
 Arg Asn Glu Lys Lys Arg His Arg Ala Val Arg Leu Ile Phe Thr Ile
 225 230 235 240
 Met Ile Val Tyr Phe Leu Phe Trp Ala Pro Tyr Asn Ile Val Leu Leu
 245 250 255
 Leu Asn Thr Phe Gln Glu Phe Phe Gly Leu Asn Asn Cys Ser Ser Ser
 260 265 270
 Asn Arg Leu Asp Gln Ala Met Gln Val Thr Glu Thr Leu Gly Met Thr
 275 280 285
 His Cys Cys Ile Asn Pro Ile Ile Tyr Ala Phe Val Gly Glu Lys Phe
 290 295 300
 Arg Asn Tyr Leu Leu Val Phe Phe Gln Lys His Ile Ala Lys Arg Phe
 305 310 315 320
 Cys Lys Cys Cys Ser Ile Phe Gln Gln Glu Ala Pro Glu Arg Ala Ser
 325 330 335
 Ser Val Tyr Thr Arg Ser Thr Gly Glu Gln Glu Ile Ser Val Gly Leu
 340 345 350

<210> 4

<211> 15

<212> DNA

<213> Homo sapien

<400> 4
gagcaatmgt atttt

15

<210> 5
<211> 15
<212> DNA
<213> Homo sapien

<400> 5
ctatgtakgc aatta

15

<210> 6
<211> 15
<212> DNA
<213> Homo sapien

<400> 6
cagggctyta tttta

15

<210> 7
<211> 15
<212> DNA
<213> Homo sapien

<400> 7
ctcatttkcc ataca

15

<210> 8
<211> 15
<212> DNA
<213> Homo sapien

<400> 8
taaagggagc aatmg

15

<210> 9
<211> 15
<212> DNA
<213> Homo sapien

<400> 9
gttattaataa tacka

15

<210> 10
<211> 15
<212> DNA
<213> Homo sapien

<400> 10
agacatctat gtakg

15

<210> 11
<211> 15
<212> DNA
<213> Homo sapien

<400> 11
ggtttttaat tgcmt

15

<210> 12
<211> 15
<212> DNA
<213> Homo sapien

<400> 12
tcttgacagg gctyt

15

<210> 13
<211> 15
<212> DNA
<213> Homo sapien

<400> 13
agcctataaaa atara

15

<210> 14
<211> 15
<212> DNA
<213> Homo sapien

<400> 14
gcagctctca tttkc

15

<210> 15

<211> 15
<212> DNA
<213> Homo sapien

<400> 15
actgactgta tggma

15

<210> 16
<211> 10
<212> DNA
<213> Homo sapien

<400> 16
agggagcaat

10

<210> 17
<211> 10
<212> DNA
<213> Homo sapien

<400> 17
attaaaatac

10

<210> 18
<211> 10
<212> DNA
<213> Homo sapien

<400> 18
catctatgta

10

<210> 19
<211> 10
<212> DNA
<213> Homo sapien

<400> 19
ttttaattgc

10

<210> 20
<211> 10
<212> DNA
<213> Homo sapien

<400> 20
tgacagggct

10

<210> 21
<211> 10
<212> DNA
<213> Homo sapien

<400> 21
ctataaaata

10

<210> 22
<211> 10
<212> DNA
<213> Homo sapien

<400> 22
gctctcattt

10

<210> 23
<211> 10
<212> DNA
<213> Homo sapien

<400> 23
gactgtatgg

10

<210> 24
<211> 9141
<212> DNA
<213> Homo sapien

<220>
<221> allele
<222> (1629)
<223> PS1: polymorphic base G or A

<220>
<221> allele
<222> (1953)
<223> PS2: polymorphic base T or C

<220>

<221> allele
<222> (1956)
<223> PS3: polymorphic base C or T

<220>
<221> allele
<222> (2002)
<223> PS4: polymorphic base G or A

<220>
<221> allele
<222> (2253)
<223> PS5: polymorphic base C or T

<220>
<221> allele
<222> (2402)
<223> PS6: polymorphic base A or C

<220>
<221> allele
<222> (3920)
<223> PS7: polymorphic base G or T

<220>
<221> allele
<222> (4246)
<223> PS8: polymorphic base T or A

<220>
<221> allele
<222> (4307)
<223> PS9: polymorphic base T or C

<220>
<221> allele
<222> (4403)
<223> PS10: polymorphic base C or T

<220>
<221> allele
<222> (4628)
<223> PS11: polymorphic base T or G

<220>
<221> allele
<222> (4750)
<223> PS12: polymorphic base G or A

<400> 24

gtatgtcttc acatcctcca atacccta atacccta gacagacaaa cagaacatgg caaagcctca 60
 gctctgcatg gtgaaagtaa gaaccagcaa ttgccacaaa cagaaatata gtgttggtcc 120
 ggcagcctcc gggggttctg cacagtgga ttaccagtga atacaaggct atctatcttt 180
 cgaaaaacca acgttgattt tatgtatttt tctataaaat tttatattaa tttatttgtt 240
 acctattttt gaactctttc aaaagcacac tttatatttc cctgcttaaa cagtcccccg 300
 aggggtgggtg cccaaaaggo tctacacttg ttatcattcc ctctccacca caggcatatt 360
 gagtaagttt gtatttgggt ttttttaaaa cctccactct acagttaaga aaactaaggc 420
 acagagcttc aataatttgg tcagagccaa gtagcagtaa tgaagctgga gggttaaacc 480
 agcagcatga ctgcagttct taatcaatgc cttttgaatt gcacatatgg gatgaactag 540
 aacattttct cgatgattcg ctgtccttgt tatgattatg ttactgagct ctgttgtagc 600
 acagacatat gtccctatat gggggtgggt tgggggtgtc ttgatcgctg ggctatttct 660
 atactgttct ggcttttccc aagcagtcac ttctttctat tctccaagca ccagcaatta 720
 gctttacctt ttcagcttct agtttgctga aactaatctg ctatagacag agactccggt 780
 gaaccaattt tattaggatt tgatcaaata aactctctct gacaaaggac tgctgaaaga 840
 gtaactaaga gtttgatgtt tactgagtg atagtatgt ctagatgctg gccgtggatg 900
 cctcatagaa tcttccaac aactcatgaa atgactactg tcattcagcc caataccag 960
 acgagaaagc tgagggttaag acaggtttca agcttggcag tctgactaca gaggccactg 1020
 gcttagcccc tgggttagtc tgcctctgta ggattggggg cagtaattt tgctgtttgg 1080
 ggtctcattt gccttcttag agatcacaa ccaaagcttt ttattctaga gccaaagtca 1140
 cggaagccca gagggcatct tgtggctcgg gagtagctct ctgctgtctt ctcagctctg 1200
 ctgacaatac ttgagatttt cagatgtcac caaccacaa gagagcttga tatgactgta 1260
 tatagtatag tcataaagaa cctgaacttg acctataact tatgtcatgt ggaaaatttc 1320
 tcatagcttc agatagatta tatctggagt gaagaatcct gccacctatg tatctggcat 1380
 agtgtgagtc ctcataaatg cttactggtt tgaaggga caaaatagt aacagagtga 1440
 aaatccccac taagatcctg ggtccagaaa aagatgggaa acctgtttag ctcaccctg 1500
 agcccatagt taaaactctt tagacaacag gtttttccg ttacagaga acaataatat 1560
 tgggtggtga gcatctgtgt ggggttggg gtgggatagg ggatacgggg agagtggaga 1620
 aaaaggggrc acagggttaa tgtgaagtcc aggatcccc tctacattta aagtgggtt 1680
 aagtgggtt taattaatag caactcttaa gataatcaga attttcttaa ccttttagcc 1740
 ttactgttga aaagccctgt gatcttgtac aaatcatttg cttcttggat agtaatttct 1800
 tttactaaaa tgtgggcttt tgactagatg aatgtaaag ttcttctagc tctgatattc 1860
 tttattcttt atattttcta acagattctg tgtagtggga tgagcagaga acaaaaacaa 1920
 aataatccag tgagaaaagc ccgtaataa acythyagac cagagatcta ttctctagct 1980
 tattttaagc tcaacttaaa argaagaact gttctctgat tcttttcgcc ttcaatacac 2040
 ttaatgattt aactccacc tcttcaaaa gaaacagcat ttcctacttt tatactgtct 2100
 atatgattga ttgacacag ccatctggcc agaagagctg agacatccgt tcccctacaa 2160
 gaaactctcc ccgtaagta acctctcagc tgcctggcct gttagttagc ttctgagatg 2220
 agtaaaagac ttacaggaa acccatagaa gayatttggc aaacaccaag tgctcataca 2280
 attatcttaa aatataatct ttaagataag gaaaggtca cagtttggaa tgagtttcag 2340
 acggttataa catcaaagat acaaaacatg attgtgagt aaagacttta aaggagcaa 2400
 tmgtatttta ataactaaca atccttacct ctcaaaagaa agatttgcag agagatgagt 2460
 cttagctgaa atcttgaaat cttatcttct gctaaggaga actaaacct ctccagttag 2520
 atgccttctg aatatgtgcc cacaagaagt tgtgtctaag tctggttctc ttttttctt 2580
 ttctccaga caagaggga gcctaaaaat ggtcaaaatt aatattaaat tacaacgcc 2640
 aaataaaatt ttctctaat atatcagtt catggcacag ttagtatata attctttatg 2700
 gtcaaaaatt aaaaatgagc ttttctaggg gcttctctca gctgcctagt ctaaggtgca 2760

gggagtttga gactcacagg gtttaataag agaaaattct cagctagagc agctgaactt 2820
 aaatagacta ggcaagacag ctggttataa gactaaacta cccagaatgc atgacattca 2880
 tctgtggtgg cagacgaaac attttttatt atattatttc ttgggtatgt atgacaactc 2940
 ttaattgtgg caactcaaac tacaacacac aacttcacag aaaatgtgag gattttacaa 3000
 ttggctgttg tcatctatga ccttccctgg gacttgggca cccggccatt tcaactctgac 3060
 tacatcatgt caccaaact ctgatggtct tgccttttaa ttctctttt gaggactgag 3120
 agggagggtg gcatggtagt taagagtga ggcttcccg attcaaaatc ggttgcttac 3180
 tagctgtgtg gctttgagca agttactcac cctctctgtg cttcaaggtc cttgtctgca 3240
 aaatgtgaaa aatatttctt gcctcataag gttgccctaa ggattaaatg aatgaatggg 3300
 tatgatgctt agaacagtga ttggcatcca gtatgtgcc tggaggcctc ttaattatta 3360
 ctggcttgct catagtgc atgtcttctg ggctaactct agcgtcaata aaaatgttaa 3420
 gactgagttg cagctgggca tgggtggctca tgcctgtaat cccagcattc taggaggctg 3480
 aggcaggagg atcgcttgag cccaggagtt cgagaccagc ctgggcaaca tagtgtgac 3540
 ttgtatctat aaaaataaac aaaattagct tgggtgtgtg gcgcctgtag tccccagcca 3600
 cttggagggg tgaggtgaga ggattgcttg agccgggat gatccaggct gcagttagcc 3660
 atgatcgtgc cactgcactc cagcctgggc gacagagtga gaccctgtct cacaacaaca 3720
 acagcaacaa aaaggctgag ctgcaccatg cttgaccag tttcttaaaa ttgttgtcaa 3780
 agcttcattc actccatggt gctatagagc acaagatttt atttggtgag atgggtgctt 3840
 catgaattcc cccaaacagag ccaagctctc catctagtgg acagggaagc tagcagcaaa 3900
 ccttcccttc actacaaaac ttcattgctt ggccaaaaag agagttaatt caatgtagac 3960
 atctatgtak gcaattaaaa acctattgat gtataaaaca gtttgcatc atggagggca 4020
 actaaataca ttctaggact ttataaaaaga tcacttttta tttatgcaca ggggtggaaca 4080
 agatggatta tcaagtgtca agtccaatct atgacatcaa ttattataca tccggagccct 4140
 gccaaaaaat caatgtgaag caaatcgag cccgcctcct gcctccgctc tactcactgg 4200
 tgttcatctt tggttttgtg ggcaacatgc tggtcacct catccwgata aactgcaaaa 4260
 ggctgaagag catgactgac atctacctgc tcaacctggc catctcygac ctgtttttcc 4320
 ttcttactgt ccccttctgg gctcactatg ctgccgccca gtgggacttt ggaaatacaa 4380
 tgtgtcaact cttgacaggg ctytatttta taggcttctt ctctggaatc ttcttcatca 4440
 tcctctgac aatcgatagg tacctggctg tcgtccatgc tgtgtttgct ttaaaagcca 4500
 ggacggtcac ctttgggtg gtgacaagt tgatcacttg ggtgggtggct gtgtttgcgt 4560
 ctctcccagg aatcatcttt accagatctc aaaaagaagg tcttcattac acctgcagct 4620
 ctcatthkcc atacagtcag tatcaattct ggaagaattt ccagacatta aagatagtca 4680
 tcttggggct ggtcctgccg ctgcttgtca tggtcacctg ctactcgga atcctaaaaa 4740
 ctctgcttcr gtgtcgaaat gagaagaaga ggcacagggc tgtgaggctt atcttcacca 4800
 tcatgattgt ttattttctc ttctgggctc cctacaacat tgtccttctc ctgaacacct 4860
 tccaggaatt ctttggcctg aataattgca gtatgtctaa caggttggaac caagctatgc 4920
 aggtgacaga gactcttggg atgacgcact gctgcatcaa ccccatcatc tatgcctttg 4980
 tccggggagaa gttcagaaac tacctcttag tcttcttcca aaagcacatt gccaaacgct 5040
 tctgcaaatg ctgttctatt ttccagcaag aggcctccga gcgagcaagc tcagtttaca 5100
 cccgatccac tggggagcag gaaatatctg tgggcttgtg acacggactc aagtgggctg 5160
 gtgaccagc cagagttgtg cacatggctt agttttcata cacagcctgg gctgggggtg 5220
 ggggtgggaga ggtctttttt aaaaggaagt tactgttata gaggttctaa gattcatcca 5280
 tttatttggc atctgtttta agtagattag atcttttaag cccatcaatt atagaaagcc 5340
 aaatcaaaat atgttgatga aaaatagcaa cctttttatc tccccctcac atgcatcaag 5400
 ttattgacaa actctccctt cactccgaaa gttccttatg tatatttaaa agaaagcctc 5460
 agagaattgc tgattcttga gtttagtgat ctgaacagaa ataccaaaat tatttcagaa 5520
 atgtacaact ttttacctag tacaaggcaa catatagggt gtaaatgtgt ttaaaacagg 5580
 tctttgtctt gctatgggga gaaaagacat gaatatgatt agtaaagaaa tgacactttt 5640

catgtgtgat ttccctcca aggtatggtt aataagtttc actgacttag aaccaggcga 5700
 gagacttggtg gcctgggaga gctggggaag cttcttaaata gagaaggaat ttgagttgga 5760
 tcactctattg ctggcaaaga cagaagcctc actgcaagca ctgcatgggc aagcttggtc 5820
 gtagaaggag acagagctgg ttgggaagac atggggagga aggacaaggc tagatcatga 5880
 agaaccttga cggcattgct ccgtctaagt catgagctga gcaggagat cctggttggt 5940
 gttgcagaag gtttactctg tgcccaaagg agggtcagga aggatgagca tttaggggca 6000
 ggagaccacc aacagccctc aggtcagggt gaggatggcc tctgctaagc tcaaggcgtg 6060
 aggatgggaa ggaggagggt attcgtaagg atgggaagga gggagggtatt cgtgcagcat 6120
 atgaggatgc agagtcagca gaactggggt ggatttggtt tggaagtgag ggtcagagag 6180
 gagtcataga gaatccctag tcttcaagca gattggagaa acccttgaag agacatcaag 6240
 cacagaagga ggaggaggag gtttaggtca agaagaagat ggattggtgt aaaaggatgg 6300
 gtctggtttg cagagcttga acacagtctc acccagactc caggctgtct ttcactgaat 6360
 gcttctgact tcatagattt ccttcccatc ccagctgaaa tactgagggg tctccaggag 6420
 gagactagat ttatgaatac acgaggtatg aggtctagga acatacttca gctcacacat 6480
 gagatctagg tgaggattga ttacctagta gtcatttcat gggttgttgg gaggattcta 6540
 tgaggcaacc acaggcagca tttagcacat actacacatt caataagcat caaactctta 6600
 gttactcatt cagggatagc actgagcaaa gcattgagca aaggggtccc atagagggtga 6660
 gggaagcctg aaaaactaag atgctgcctg cccagtgcac acaagtgtag gtatcatttt 6720
 ctgcatttaa ccgtcaatag gcaaaggggg gaagggacat attcatttgg aaataagctg 6780
 ccttgagcct taaaaccac aaaagtacaa tttaccagcc tccgtatttc agactgaatg 6840
 ggggtggggg gggcgccctt ggtacttatt ccagatgcct tctccagaca aaccagaagc 6900
 aacagaaaaa atcgtctctc cctccctttg aaatgaatat accccttagt gtttggttat 6960
 attcatttca aagggagaga gagaggtttt tttctgttct ttctcatatg attgtgcaca 7020
 tacttgagac tgttttgaat ttgggggatg gctaaaacca tcatagtaca ggtaagggtga 7080
 gggaatagta agtggtgaga actactcagg gaatgaagggt gtcagaataa taagagggtgc 7140
 tactgacttt ctgagcctct gaatatgaac ggtgagcatt gtggctgtca gcaggaagca 7200
 acgaagggaa atgtctttcc ttttgctctt aagttgtgga gagtgaaca gtagcatagg 7260
 accctaccct ctgggccaag tcaaagacat tctgacatct tagtatttgc atattcttat 7320
 gtatgtgaaa gttacaatt gcttgaaaga aaatatgcat ctaataaaaa acaccttcta 7380
 aaataattca ttatattctt gctctttcag tcaagtgtac atttagagaa tagcacataa 7440
 aactgccaga gcattttata agcagctgtt ttcttcttga gtgtgtgtgc atgtgtgtgt 7500
 gatgtatata aagagagaga taattgtatt tttgtatttt cttttaataa atttttaaaa 7560
 ttgacccttt tctgagaca aattgccaga atagtttgta tttagagatg gtacctctaa 7620
 gagtaagggt gctggttgc gagcaattga cttgaaaact tttaaaattc aaattttaat 7680
 tccactactc aaaagaattg ccatgtttta aaaaagagaa ttggtgccat aagttagttg 7740
 tctatgtttg aaaatgaaga agatatgcaa cgtcatggcc tggtcactta cccgcagccc 7800
 tgagttgtag gcacatcata tgtgagaatg aggatgcttt tctttcattt aaaatccctc 7860
 cccaaaactt ggctctaatt gcagtcatga caatcatgta catttggtat tatgtgcacg 7920
 agtctcttac cctgagagag gacagggtgt acagggtggag gggaccgtc tgggtcacgt 7980
 tcacattttg aacatgctgg ttttcagtca ctgcacactc atctcccagc gcaggatcatg 8040
 ggcagcagat gcaaaagctg cccgtggtcc tatttgaggg tgcatgaaat gagcagaaga 8100
 cagaatagct tgatctgact agaagggcag cttgtcccta ccaagacttg aaggattgcc 8160
 tttcatctgt tagggtaaaa ggtagaatga accaaggaag ggcaggaggg ggctgggggt 8220
 agggtagaag gaaggggcca tggagaagg agatccatcc cataggagga aggcagtgcg 8280
 gcaggagggt ttgaaggat cagcttttgt ggctgacata catgcagtca tgtcaattgc 8340
 tctgttttcc ttttccatct tattaatgt cttccaatgt tagcacgaag aaaagctatt 8400
 tgcagtgtt ccagccttc cagagcccgt cccattacc tcccaggcc catgccttta 8460
 ctccttgag tttcaactca cgacctcag gatctgactt tattcacaa ctctgggggtg 8520

aacgtacctt ctgtctccac ocagaggtct ctatcaaaga ggagattgca tgccatggat 8580
aaagtcaaag tagaggtgac tgtccttagg aagagtaatg tgaaaattca taaactggga. 8640
ttctgtttac attttgtact ccaggggttc ttagtttaaa tcgctctgaa taaattaaga 8700
tgcaatggca tttcaactgt tatgattaaa ttacaaaac atttattttc tatcacgggg 8760
agagatagag ctccaaatgc aaacataact gctcaagtgt taacacttat aatgaaaaca 8820
taagaattac caccaactac cctgggggct agaagcagaa atgtgaacca gaaaacaaat 8880
catgaacttt cctttttttt tttgagatgg agtctcgctc tgttgcccag gctggagtgc 8940
aatggtgcga tctcggtcga ctgcaaccac tgcctcccgg gttcaagcaa ttctcctgcc 9000
tcagcctcct gagtagctgg gactacaggc atgcaccacc acgcctgggt aattttttgt 9060
attttttagta gagacagggg ttcaccgtat tagccaggat gctcttgatc tcctgacctc 9120
gtgatctgcc cgcctcggcc t 9141